Journal of Machine Learning Research () Submitted 02/01; Published

A Convergent Online Single Time Scale Actor Critic Algorithm

Dotan Di Castro DOT@TX.TECHNION.AC.IL
Ron Meir RMEIR@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion, Haifa 32000, Israel

Editor:

arXiv:0909.2934v1 [cs.LG] 16 Sep 2009

Abstract

Actor-Critic based approaches were among the first to address reinforcement learning in a general setting. Recently, these algorithms have gained renewed interest due to their generality, good convergence properties, and possible biological relevance. In this paper, we introduce an online temporal difference based actor-critic algorithm which is proved to converge to a neighborhood of a local maximum of the average reward. Linear function approximation is used by the critic in order estimate the value function, and the temporal difference signal, which is passed from the critic to the actor. The main distinguishing feature of the present convergence proof is that both the actor and the critic operate on a similar time scale, while in most current convergence proofs they are required to have very different time scales in order to converge. Moreover, the same temporal difference signal is used to update the parameters of both the actor and the critic. A limitation of the proposed approach, compared to results available for two time scale convergence, is that convergence is guaranteed only to a neighborhood of an optimal value, rather to an optimal value itself. The single time scale and identical temporal difference signal used by the actor and the critic, may provide a step towards constructing more biologically realistic models of reinforcement learning in the brain.

1. Introduction

In Reinforcement Learning (RL) an agent attempts to improve its performance over time at a given task, based on continual interaction with the (usually unknown) environment (Bertsekas and Tsitsiklis (1996); Sutton and Barto (1998)). Formally, it is the problem of mapping situations to actions in order to maximize a given average reward signal. The interaction between the agent and the environment is modeled mathematically as a Markov Decision Process (MDP). Approaches based on a direct interaction with the environment, are referred to as *simulation based algorithms*, and will form the major focus of this paper.

A well known subclass of RL approaches consists of the so called actor-critic (AC) algorithms (e.g., Sutton and Barto (1998)), where the agent is divided into two components, an actor and a critic. The critic functions as a value estimator, whereas the actor attempts to select actions based on the value estimated by the critic. These two components solve their own problems separately but interactively. Many methods for solving the critic's value estimation problem, for a *fixed* policy, have been proposed, but, arguably, the most widely used is *temporal difference* (TD) learning. TD learning was demonstrated to accelerate convergence by trading bias for variance effectively Singh and Dayan (1998), and is often used as a component of AC algorithms.

In general, policy selection may be randomized. When facing problems with a large number of states or actions (or even continuous state-action problems), effective policy selection may suffer from several problems, such as slow convergence rate or an inefficient representation of the policy. A possible approach to policy learning is the so-called *policy gradient method* (Baxter and Bartlett

© D. Di Castro and R. Meir 2009.

(2001); Cao (2007); Cao and Chen (1997); Konda and Tsitsiklis (2003); Marbach and Tsitsiklis (1998)). Instead of maintaining a separate estimate for the value for each state (or state-action pair), the agent maintains a parametrized policy function. The policy function is taken to be a differentiable function of a parameter vector and of the state. Given the performance measure, depending on the agent's policy parameters, these parameters are updated using a sampling-based estimate of the gradient of the average reward. While such approaches can be proved to converge under certain conditions (e.g., Baxter and Bartlett (2001)), they often lead to slow convergence, due to very high variance. A more general approach based on sensitivity analysis, which includes policy gradient methods as well as non-parametric average reward functions, has been discussed in depth in the recent manuscript by Cao (2007).

Several AC algorithms with associated convergence proofs have been proposed recently (a short review is given in section 2.2). As far as we are aware, all the convergence results for these algorithms are based on two time scales, specifically, the actor is assumed to update its internal parameters on a much slower time scale than the one used by the critic. The intuitive reason for this time scale separation is clear, since the actor improves its policy based on the critic's estimates. It can be expected that rapid change of the policy parameters may not allow the critic to effectively evaluate the value function, which may lead to instability when used by the actor in order to re-update its parameters.

The objective of this paper is to propose an online AC algorithm and establish its convergence under conditions which do *not* require the separation into two time scales. There is clear theoretical motivation for such an approach,as it can potentially lead to faster convergence rates, although this is not a an issue we stress in this work. In fact, our motivation for the current direction was based on the possible relevance of AC algorithms in a biological context (e.g, Daw et al. (2006)), where it would be difficult to justify two very different time scales operating within the same anatomical structure. We refer the reader to DiCastro et al. (2008) for some preliminary ideas and references related to these issues. Given the weaker conditions assumed on the time scales, our convergence result is, not surprisingly, somewhat weaker than that provided recently in (e.g., Bhatnagar et al. (2008a,b)), as we are not ensured to converge to a local optimum, but only to a neighborhood of such an optimum. Nevertheless, it is shown that the neighborhood size can be algorithmically controlled. Further comparative discussion can be found in section 2.

This paper is organized as follows. In section 2 we briefly recapitulate current AC algorithms for which convergence proofs are available. In section 3, we formally introduce the problem setup. We begin section 4 by relating the TD signal to the gradient of the average reward, and then move on to motivate and derive the main AC algorithm, concluding the section with a convergence proof. A comparative discussion of the main features of our approach is presented in section 5, followed by some simulation results in section 6. Finally, in section 7, we discuss the results and point out possible future work. In order to facilitate the readability of the paper, we have relegated all technical proofs to appendices.

2. Previous Work

In this section we briefly review some previous work in RL which bears direct relevance to our work. While many AC algorithms have been introduced over the years, we focus only on those for which a convergence proof is available, since the main focus of this work is on convergence issues, rather than on establishing the most practically effective algorithms (see, for example, Peters and Schaal (2008), for promising applications of AC algorithms in a robotic setting).

2.1 Direct policy gradient algorithms

Direct policy gradient algorithms, employing agents which consist of an actor only, typically estimate a noisy gradient of the average reward, and are relatively close in their characteristics to AC

algorithms. The main difference from the latter is that the agent does not maintain a separate value estimator for each state, but rather interacts with the environment directly, and in a sense maintains its value estimate implicitly through a mapping which signifies which path the agent should take in order to maximize its average reward per stage.

Marbach and Tsitsiklis (1998) suggested an algorithm for non-discounted environments. The gradient estimate is based on an estimate of the state values which the actor estimates while interacting with the environment. If the actor returns to a sequence of previously visited states, it re-estimates the states value, not taking into account its previous visits. This approach often results in large estimation variance.

Baxter and Bartlett (2001) proposed an online algorithm for partially observable MDPs. In this algorithm, the agent estimates the expected average reward for the non-discounted problems through an estimate of the value function of a related discounted problem. It was shown that when the discount factor approaches 1, the related discounted problem approximates the average reward per stage. Similar to the algorithms in (Marbach and Tsitsiklis (1998)), it suffers from relatively large estimation variance. In (Baxter et al. (2004)), a method was proposed for coping with the large variance by adding a baseline to the value function estimation.

2.2 Actor Critic Algorithms

As stated in section 1, the convergence proofs of which we are aware for AC algorithms are based on two time scale stochastic approximation (Borkar (1997)), where the actor is assumed to operate on a time scale which is much slower than that used by the critic.

Konda and Borkar (1999) suggested a set of AC algorithms. In two of their algorithms (Algorithms 3 and 6), parametrized policy based actors were used while the critic was based on a lookup table. Those algorithms and their convergence proofs were specific to the Gibbs policy function in the actor.

As far as we are aware, Konda and Tsitsiklis (2003) provided the first convergence proof for an AC algorithm based on function approximation. The information passed from the critic to the actor is the critic's action-value function, and the critic's basis functions, which are explicitly used by the actor. They provided a convergence proof of their TD(λ) algorithm where λ approaches 1. A drawback of the algorithm is that the actor and the critic must share the information regarding the actor's parameters. This detailed information sharing is a clear handicap in a biological context, which was one of the driving forces for the present work.

Finally, Bhatnagar et al. (2008a,b) recently proposed an AC algorithm which closely resembles our proposed algorithm, and which was developed independently of ours. In this work the actor uses a parametrized policy function while the critic uses a function approximation for the state evaluation. The critic passes to the actor the TD(0) signal and based on it the actor estimates the average reward gradient. A detailed comparison will be provided in section 5. As pointed out in Bhatnagar et al. (2008a,b), their work is the first to provide a convergence proof for an AC algorithm incorporating bootstrapping Sutton and Barto (1998), where bootstrapping refers to a situation where estimates are updated based on other estimates, rather than on direct measurements (as in Monte Carlo approaches). This feature applies to our work as well. We also note that Bhatnagar et al. (2008a,b) extend their approach to the so-called natural gradient estimator, which has been shown to lead to improved convergence in supervised learning as well as RL. The present study focuses on the standard gradient estimate, leaving the extension to natural gradients to future work.

3. The Problem Setup

In this section we describe the formal problem setup, and present a sequence of assumptions and lemmas which will be used in order to prove convergence of Algorithm 1 in section 4. These assump-

tions and lemmas mainly concern the properties of the controlled Markov chain, which represents the environment, and the properties of the actor's parametrized policy function.

3.1 The Dynamics of the Environment and of the Actor

We consider an agent, composed of an actor and a critic, interacting with an environment. We model the environment as a *Markov Decision Process* (MDP) Puterman (1994) in discrete time with a finite state set $\mathcal{X}$ and an action set $\mathcal{U}$, which may be uncountable. We denote by $|\mathcal{X}|$ the size of the set $\mathcal{X}$. Each selected action $u \in U$ determines a stochastic matrix $P(u) = [P(y|x,u)]_{x,y\in\mathcal{X}}$ where $P(y|x,u)$ is the transition probability from a state $x \in \mathcal{X}$ to a state $y \in \mathcal{X}$ given the control u. For each state $x \in \mathcal{X}$ the agent receives a corresponding reward $r(x)$, which may be deterministic or random. In the present study we assume for simplicity that the reward is deterministic, a benign assumption which can be easily generalized.

Assumption 3.1 *The rewards, $\{r(x)\}_{x\in\mathcal{X}}$, are uniformly bounded by a finite constant B_r.*

The actor maintains a *parametrized policy function*. A parametrized policy function is a conditional probability function, denoted by $\mu(u|x,\theta)$, which maps an observation $x \in \mathcal{X}$ into a control $u \in \mathcal{U}$ given a parameter $\theta \in \mathbb{R}^K$. The agent's goal is to adjust the parameter θ in order to attain maximum average reward over time. For each θ, we have a Markov Chain (MC) induced by $P(y|x,u)$ and $\mu(u|x,\theta)$. The state transitions of the MC are obtained by first generating an action u according to $\mu(u|x,\theta)$, and then generating the next state according to $\{P(y|x,u)\}_{x,y\in\mathcal{X}}$. Thus, the MC has a transition matrix $P(\theta) = [P(y|x,\theta)]_{x,y\in\mathcal{X}}$ which is given by

$$P(y|x,\theta) = \int_{\mathcal{U}} P(y|x,u)d\mu(u|x,\theta). \tag{1}$$

We denote the space of these transition probabilities by $\mathcal{P} = \{P(\theta)|\theta \in \mathbb{R}^K\}$, and its closure by $\bar{\mathcal{P}}$. The following assumption is needed in the sequel in order to prove the main results (see Brémaud (1999) for definitions).

Assumption 3.2 *Each MC, $P(\theta) \in \bar{\mathcal{P}}$, is aperiodic, recurrent, and irreducible.*

As a result of Assumption 3.2, we have the following lemma regarding the stationary distribution and a common recurrent state.

Lemma 3.3 *Under Assumption 3.2 we have:*

1. Each MC, $P(\theta) \in \bar{\mathcal{P}}$, has a unique stationary distribution, denoted by $\pi(\theta)$, satisfying $\pi(\theta)'P(\theta) = \pi(\theta)'$.
2. There exists a state, denoted by x^*, which is recurrent for all $P(\theta) \in \bar{\mathcal{P}}$.

Proof For the first part see Corollary 4.1 in (Gallager, 1995). The second part follows trivially from Assumption 3.2. ■

The next technical assumption states that the first and second derivatives of the parametrized policy function are bounded, and is needed to prove Lemma 3.6 below.

Assumption 3.4 *The conditional probability function $\mu(u|x,\theta)$ is twice differentiable. Moreover, there exist positive constants, B_{μ_1} and B_{μ_2}, such that for all $x \in \mathcal{X}$, $u \in \mathcal{U}$, $\theta \in \mathbb{R}^K$ and $k_1 \geq 1$, $k_2 \leq K$ we have*

$$\left|\frac{\partial\mu(u|x,\theta)}{\partial\theta_k}\right| \leq B_{\mu_1}, \quad \left|\frac{\partial^2\mu(u|x,\theta)}{\partial\theta_{k_1}\partial\theta_{k_2}}\right| \leq B_{\mu_2}.$$

A notational comment concerning bounds *Throughout the paper we denote upper bounds on*

different variables by the letter B, with a subscript corresponding to the variable itself. An additional numerical subscript, 1 or 2, denotes a bound on the first or second derivative of the variable. For example, B_f, B_{f_1}, and B_{f_2} denote the bounds on the function f and its first and second derivatives respectively.

3.2 Performance Measures

Next, we define a performance measure for an agent in an environment. The *average reward per stage* of an agent which traverses a MC starting from an initial state $x \in \mathcal{X}$ is defined by

$$J(x,\theta) \triangleq \lim_{T\to\infty} \mathrm{E}\left[\frac{1}{T}\sum_{n=0}^{T-1} r(x_n)\middle| x_0 = x, \theta\right],$$

where $\mathrm{E}[\cdot|\theta]$ denotes the expectation under the probability measure $P(\theta)$, and x_n is the state at time n. The agent's goal is to find $\theta \in \mathbb{R}^K$ which maximizes $J(x,\theta)$. The following lemma shows that under Assumption 3.2, the average reward per stage does not depend on the initial state (Bertsekas (2006), vol. II, section 4.1).

Lemma 3.5 *Under Assumption 3.2 and based on Lemma 3.3, the average reward per stage, $J(x,\theta)$, is independent of the starting state, is denoted by $\eta(\theta)$, and satisfies $\eta(\theta) = \pi(\theta)'r$.*

Based on Lemma 3.5, the agent's goal is to find a parameter vector θ, which maximizes the average reward per stage $\eta(\theta)$. In the sequel we show how this maximization can be performed by optimizing $\eta(\theta)$, using $\nabla_\theta \eta(\theta)$. A consequence of Assumption 3.4 and the definition of $\eta(\theta)$ is the following lemma.

Lemma 3.6

1. For each $x, y \in \mathcal{X}$, $1 \le i, j \le K$, and $\theta \in \mathbb{R}^K$, the functions $\partial P(y|x,\theta)/\partial\theta_i$ and $\partial^2 P(y|x,\theta)/\partial\theta_i\partial\theta_j$ are uniformly bounded by B_{P_1} and B_{P_2} respectively.
   (a) For each $x \in \mathcal{X}$, $1 \le i, j \le K$, and $\theta \in \mathbb{R}^K$, the functions $\partial\pi(x|\theta)/\partial\theta_i$ and $\partial^2\pi(x|\theta)/\partial\theta_i\partial\theta_j$ are uniformly bounded by , $B_{\pi_1}$and B_{π_2} respectively.
   (b) For all $1 \le i, j \le K$, and $\theta \in \mathbb{R}^K$, the functions $\eta(\theta)$, $\partial\eta(\theta)/\partial\theta_i$ and $\partial^2\pi(x|\theta)/\partial\theta_i\partial\theta_j$ are uniformly bounded by , B_η, B_{η_1} and B_{η_2} respectively.
   (c) For all $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^K$, there exists a constant $b_\pi > 0$ such that $\pi(x|\theta) \ge b_\pi$.

The proof is technical and is given in Appendix A.1. For later use, we define the random variable T, which denotes the first return time to the recurrent state x^*. Formally,

$$T \triangleq \min\{k > 0 | x_0 = x^*, x_k = x^*\}. \tag{2}$$

It is easy to show that under Assumption 3.2, the average reward per stage can be expressed by

$$\eta(\theta) = \lim_{T\to\infty} \mathrm{E}\left[\frac{1}{T}\sum_{n=0}^{T-1} r(x_n)\middle| x_0 = x^*, \theta\right]. \tag{3}$$

Next, we define the *differential value function* of state $x \in \mathcal{X}$ which represents the average differential reward the agent receives upon starting from a state x and reaching the recurrent state x^* for the first time. Mathematically,

$$h(x,\theta) \triangleq \mathrm{E}\left[\sum_{n=0}^{T-1}(r(x_n) - \eta(\theta))\middle| x_0 = x, \theta\right]. \tag{4}$$

Abusing notation slightly, we denote $h(\theta) \triangleq (h(x_1,\theta),\ldots,h(x_{|\mathcal{X}|},\theta)) \in \mathbb{R}^{|\mathcal{X}|}$. For each $\theta \in \mathbb{R}^K$ and $x \in \mathcal{X}$, $h(x,\theta)$, $r(x)$, and $\eta(\theta)$ satisfy Poisson's equation (see Theorem 7.4.1 in (Bertsekas (2006))), i.e.,

$$h(x,\theta) = r(x) - \eta(\theta) + \sum_{y\in\mathcal{X}} P(y|x,\theta)h(y,\theta). \tag{5}$$

Based on the differential value we define the *temporal difference* (TD) between the states $x \in \mathcal{X}$ and $y \in \mathcal{X}$ (see Bertsekas and Tsitsiklis (1996), Sutton and Barto (1998)),

$$d(x,y,\theta) \triangleq r(x) - \eta(\theta) + h(y,\theta) - h(x,\theta). \tag{6}$$

According to common wisdom, the TD is interpreted as a prediction error. The next lemma states the boundedness of $h(x,\theta)$ and its derivatives. The proof is given in Appendix A.2.

Lemma 3.7

1. The differential value function, $h(x,\theta)$, is bounded and has bounded first and second derivative. Mathematically, for all $x \in \mathcal{X}$, $1 \le i,j \le K$, and for all $\theta \in \mathbb{R}^K$ we have

$$|h(x,\theta)| \le B_h, \quad \left|\frac{\partial h(x,\theta)}{\partial \theta_i}\right| \le B_{h_1}, \quad \left|\frac{\partial^2 h(x,\theta)}{\partial \theta_i \partial \theta_j}\right| \le B_{h_2}.$$

   (a) There exists a constant B_D such that or all $\theta \in \mathbb{R}^K$ we have $|d(x,y,\theta)| \le B_D$, where $B_D = 2\left(B_r + B_h\right)$.

3.3 The Critic's Dynamics

The critic maintains an estimate of the environmental state values. It does so by maintaining a parametrized function which approximates $h(x,\theta)$, and is denoted by $\tilde{h}(x,w)$. The function $\tilde{h}(x,w)$ is a function of the state $x \in \mathcal{X}$ and a parameter $w \in \mathbb{R}^L$. We note that $h(x,\theta)$ is a function of θ, and is induced by the actor policy $\mu(u|x,\theta)$, while $\tilde{h}(x,w)$ is a function of w. Thus, the critic's objective is to find the parameter w which yields the best approximation of $h(\theta) = (h(x_1,\theta),\ldots,h(x_{|\mathcal{X}|},\theta))$, in a sense to be defined later. We denote this optimal vector by $w^*(\theta)$. An illustration of the interplay between the actor, critic, and the environment is given in Figure 1.

4. A Single Time Scale Actor Critic Algorithm with Linear Function Approximation

In this section, we present a version of an AC algorithm, along with its convergence proof. The core of the algorithm is based on (7) below, where the actor's estimate of $\nabla_\theta \eta(\theta)$ is based on the critic's estimate of the TD signal $d(x,y,\theta)$. The algorithm is composed of three iterates, one for the actor and two for the critic. The actor maintains the iterate of the parameter vector θ corresponding to the policy $\mu(u|x,\theta)$, where its objective is to find the optimal value of θ, denoted by θ^*, which maximizes $\eta(\theta)$. The critic maintains the other two iterates. One iterate is used for estimating the average reward per stage, $\eta(\theta)$, where its estimate is denoted by $\tilde{\eta}$. The critic's second iterate maintains a parameter vector, denoted by $w \in \mathbb{R}^L$, which is used for the differential value estimate using a function approximator, denoted by $\tilde{h}(w)$. For each $\theta \in \mathbb{R}^K$, there exists a $w^*(\theta)$ which, under the policy induced by θ, is the optimal w for estimating $\tilde{\eta}(w)$. The critic's objective is to find the optimal $\tilde{\eta}$ and w.

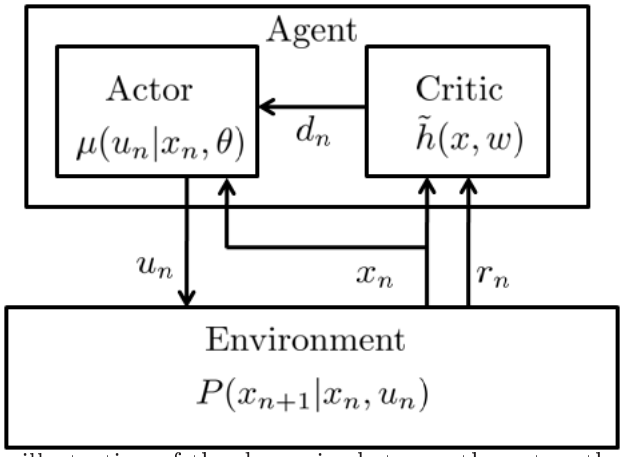


Figure 1: A schematic illustration of the dynamics between the actor, the critic, and the environment. The actor chooses an action, u_n, according to the parametrized policy $\mu(u|x,\theta)$. As a result, the environment proceeds to the next state according to the transition probability $P(x_{n+1}|x_n,u_n)$ and provides a reward. Using the TD signal, the critic improves its estimation for the environment state values while the actor improves its policy.

4.1 Using the TD Signal to Estimate the Gradient of the Average Reward

We begin with a theorem which serves as the foundation for the policy gradient algorithm described in Section 4. The theorem relates the gradient of the average reward per stage, $\eta(\theta)$, to the TD signal. It was proved in (Bhatnagar et al. (2008a)), and is similar in its structure to other theorems which connect $\eta(\theta)$ to the Q-value (Konda and Tsitsiklis (2003)), and to the differential value function (Cao (2007); Marbach and Tsitsiklis (1998)).

We start with a definition of the *likelihood ratio derivative*

$$\psi(x,u,\theta) \triangleq \frac{\nabla_\theta \mu(u|x,\theta)}{\mu(u|x,\theta)},$$

where the gradient ∇_θ is w.r.t. θ, and $\psi(x,u,\theta) \in \mathbb{R}^K$. The following assumption states that $\psi(x,u,\theta)$ is bounded, and will be used to prove the convergence of algorithm 1.

Assumption 4.1 *For all $x \in \mathcal{X}$, $u \in \mathcal{U}$, and $\theta \in \mathbb{R}^K$, there exists a positive constant, B_ψ, such that*

$$\|\psi(x,u,\theta)\|_2 \leq B_\psi < \infty,$$

where $\|\cdot\|_2$ is the Euclidean L_2 norm.

Based on this, we present the following theorem which relates the gradient of $\eta(\theta)$ to the TD signal. For completeness, we supply a (straightforward) proof in Appendix B.

Theorem 4.2 *For any arbitrary function $f(x)$, the gradient w.r.t. θ of the average reward per stage can be expressed by*

$$\nabla_\theta \eta(\theta) = \sum_{x,y \in \mathcal{X}} P(x,u,y,\theta)\psi(x,u,\theta)d(x,y,\theta), \tag{7}$$

where $P(x,u,y,\theta)$ is the probability $\Pr(x_n = x, u_n = u, x_{n+1} = y)$ subject to the policy parameter θ.

4.2 The updates performed by the critic and the actor

We note that the following derivation regarding the critic is similar in some respects to the derivation in section 6.3.3 of Bertsekas and Tsitsiklis (1996) and of Tsitsiklis and Roy (1997). We define the following quadratic target function used to evaluate the critic's performance in assessing the differential value $h(\theta)$,

$$I(w,\theta) \triangleq \frac{1}{2}\sum_{x\in\mathcal{X}} \pi(x|\theta)\left(\tilde{h}(x,w)-h(x,\theta)\right)^2 . \tag{8}$$

The probabilities $\{\pi(x|\theta)\}_{x\in\mathcal{X}}$ are used in order to provide the proportional weight to the state estimates, according to the relative number of visits of the agent to the different states.

Limiting ourselves to the class of linear function approximations in the critic, we consider the following function for the differential value function

$$\tilde{h}(x,w) = \phi(x)'w, \tag{9}$$

where $\phi(x) \in \mathbb{R}^L$. We define $\Phi \in \mathbb{R}^{|\mathcal{X}|\times L}$ to be the matrix

$$\Phi \triangleq \begin{pmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_L(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_L(x_2) \\ \vdots & \vdots & & \vdots \\ \phi_1(x_{|\mathcal{X}|}) & \phi_2(x_{|\mathcal{X}|}) & \dots & \phi_L(x_{|\mathcal{X}|}) \end{pmatrix},$$

where $\phi(\cdot)$ is a column vector. Therefore, we can express (9) in vector form as

$$\tilde{h}(w) = \Phi w, \tag{10}$$

where, abusing notation slightly, we set $\tilde{h}(w) = \left(\tilde{h}(x_1,w),\dots,\tilde{h}(x_{|\mathcal{X}|},w)\right)'$.

We wish to express (8), and the approximation process, in an appropriate Hilbert space. Define the matrix $\Pi(\theta)$ to be a diagonal matrix $\Pi(\theta) \triangleq \mathrm{diag}(\pi(\theta))$. Thus, (8) can be expressed as

$$I(w,\theta) = \frac{1}{2}\left\|\Pi(\theta)^{\frac{1}{2}}\left(h(\theta)-\Phi w\right)\right\|_2^2 \triangleq \frac{1}{2}\left\|h(\theta)-\Phi w\right\|_{\Pi(\theta)}^2 . \tag{11}$$

In the sequel, we will need the following technical assumption.

Assumption 4.3

1. The columns of the matrix Φ are independent, i.e., they form a basis of dimension L.
   (a) The norms of the column vectors of the matrix Φ are bounded above by 1, i.e., $\|\phi_k\|_2 \le 1$ for $1 \le k \le L$.

The parameter $w^*(\theta)$, which optimizes (11), can be directly computed, but involves inverting a matrix. Thus, in order to find the right estimate for $\tilde{h}(w)$, the following *gradient descent* (Bertsekas and Tsitsiklis (1996)) algorithm is suggested,

$$w_{n+1} = w_n - \gamma_n \nabla_w I(w_n,\theta), \tag{12}$$

where $\{\gamma_n\}_{n=1}^{\infty}$ is a positive series satisfying the following assumption, which will be used in proving the convergence of Algorithm 1.

Assumption 4.4 *The positive series $\{\gamma_n\}_{n=1}^{\infty}$ satisfies*

$$\sum_{n=1}^{\infty}\gamma_n = \infty, \quad \sum_{n=1}^{\infty}\gamma_n^2 < \infty. \tag{13}$$

Writing the term $\nabla_w I(w_n)$ explicitly yields

$$\nabla_w I(w_n) = \Phi'\Pi(\theta)\Phi w_n - \Phi'\Pi(\theta)h(\theta). \tag{14}$$

For each $\theta \in \mathbb{R}^K$, the value $w^*(\theta)$ is given by setting $\nabla_w I(w,\theta) = 0$, i.e.,

$$w^*(\theta) = \left(\Phi'\Pi(\theta)\Phi\right)^{-1}\Phi'\Pi(\theta)h(\theta). \tag{15}$$

Note that Bertsekas and Tsitsiklis (1996) prove that the matrix $\left(\Phi'\Pi(\theta)\Phi\right)^{-1}\Phi'\Pi(\theta)$ is a projection operator into the space spanned by Φw, with respect to the norm $\|\cdot\|_{\Pi(\theta)}$. Thus, the explicit gradient descent procedure (12) is

$$w_{n+1} = w_n - \gamma_n \Phi'\Pi\left(\theta\right)\left(\Phi w_n - h(\theta)\right). \tag{16}$$

Using the basis Φ, in order to approximates $h\left(\theta\right)$, yields an approximation error defined by

$$\epsilon_{\text{app}}\left(\theta\right) \triangleq \inf_{w\in\mathbb{R}^L}\left\|h\left(\theta\right) - \Phi w\right\|_{\pi(\theta)} = \left\|h\left(\theta\right) - \Phi w^*\left(\theta\right)\right\|_{\pi(\theta)}.$$

We can bound this error by

$$\epsilon_{\text{app}} \triangleq \sup_{\theta\in\mathbb{R}^K}\epsilon_{\text{app}}\left(\theta\right). \tag{17}$$

The agent cannot access $h(x,\theta)$ directly. Instead, it can interact with the environment in order to estimate $h(x,\theta)$. We denote by $\hat{h}_n(x)$ the estimate of $h(x,\theta)$ at time step n, thus (16) becomes

$$w_{n+1} = w_n + \gamma_n \Phi'\Pi(\theta)\left(\hat{h}_n - \Phi w_n\right). \tag{18}$$

This procedure is termed *stochastic gradient descent* (Bertsekas and Tsitsiklis (1996)).

There exist several estimators for $\hat{h}_n$. One sound method, which performs well in practical problems (see Tesauro (1995)), is the TD(λ) method (see section 5.3.2 and 6.3.3 in Bertsekas and Tsitsiklis (1996), or Chapter 6 in Sutton and Barto (1998)), where the parameter λ satisfies $0 \leq \lambda \leq 1$. This method devises an estimator which is based on previous estimates of $h\left(w\right)$, i.e., w_n, and is based also on the environmental reward $r\left(x_n\right)$. This idea is a type of *a bootstrapping* algorithm, i.e., using existing estimates and new information in order to build more accurate estimates (see Sutton and Barto (1998), Section 6.1).

The TD(λ) estimator for $\hat{h}_{n+1}$ is

$$\hat{h}_{n+1}\left(x_n\right) \quad = \quad (1-\lambda)\sum_{k=0}^{\infty}\lambda^k \hat{h}_{n+1}^{(k)}\left(x_n\right), \tag{19}$$

where the *k-steps predictor* is defined by

$$\hat{h}_{n+1}^{(k)}\left(x_n\right) = \left(\sum_{m=0}^{k} r\left(x_{n+m}\right) + \hat{h}_n\left(x_{n+k+1}\right)\right).$$

The idea of bootstrapping is apparent in (19): the predictor for the differential value of the state x_n at the $(n+1)$-Th time step, is based partially on the previous estimates through $\hat{h}_n\left(x_{n+k+1}\right)$, and partially on new information, i.e., the reward $r\left(x_{n+m}\right)$. In addition, the parameter λ gives an exponential weighting for the different k-step predictors. Thus, choosing the right λ can yield better estimators.

For the discounted setting, it was proved by Bertsekas and Tsitsiklis (1996) (p. 295) that an algorithm which implements the TD(λ) estimator (19) online and converges to the right value is the following one

$$\begin{aligned} w_{n+1} &= w_n + \gamma_n d_n e_n, \\ e_n &= \alpha\lambda e_{n-1} + \phi\left(x_n\right), \end{aligned} \tag{20}$$

where d_n is the temporal difference between the n-th and the $(n+1)$-th cycle, and e_n is the so-called *eligibility trace* (see Sections 5.3.3 and 6.3.3 inBertsekas and Tsitsiklis (1996) or Chapter 7 in Sutton and Barto (1998)), and the parameter α is the discount factor. The eligibility trace is an auxiliary variable, which is used in order to implement the idea of (19) as an online algorithm. As the name implies, the eligibility variable measures how eligible is the TD variable, d_n, in (20).

In our setting, the non-discounted case, the analogous equations for the critic, are

$$\begin{aligned} w_{n+1} &= w_n + \gamma_n \tilde{d}\left(x_n, x_{n+1}, w_n\right) e_n \\ \tilde{d}\left(x_n, x_{n+1}, w_n\right) &= r(x_n) - \tilde{\eta}_m + \tilde{h}(x_{n+1}, w_m) - \tilde{h}(x_n, w_m) \\ e_n &= \lambda e_{n-1} + \phi\left(x_n\right). \end{aligned} \tag{21}$$

The actor's iterate is motivated by Theorem 4.2. Similarly to the critic, the actor executes a stochastic gradient ascent step in order to fior with a parametrized policy $\mu(u|x,\theta)$ satisfying Assumptions 3.4 and 4.1.

- A critic with

nd a local maximum of the average reward per stage $\eta(\theta)$. Therefore,

$$\theta_{n+1} = \theta_n + \gamma_n \psi(x_n, u_n, \theta_n)\tilde{d}_n(x_n, x_{n+1}, w_n). \tag{22}$$

A summary of the algorithm is presented in Algorithm 1.

4.3 Convergence Proof for the AC Algorithm

In the remainder of this section, we state the main theorems related to the convergence of Algorithm 1. We present a sketch of the proof in this section, where the technical details are relegated to Appendices C and D. The proof is divided into two stages. In the first stage we relate the stochastic approximation to a set of ordinary differential equations (ODE). In the second stage, we find conditions under which the ODE system converges to a neighborhood of the optimal $\eta(\theta)$.

The ODE approach is a widely used method in the theory of stochastic approximation for investigating the asymptotic behavior of stochastic iterates, such as (23)-(25). The key idea of the technique is that the iterate can be decomposed into a mean function and a noise term, such as a martingale difference noise. As the iterates advance, the effect of the noise weakens due to repeated averaging. Moreover, since the step size of the iterate decreases (e.g., γ_n in (23)-(25)), one can show that asymptotically an interpolation of the iterates converges to a continuous solution of the ODE. Thus, the first part of the convergence proof is to find the ODE system which describes the asymptotic behavior of Algorithm 1. This ODE will be presented in Theorem 4.6. In the second part we use ideas from the theory of Lyapunov functions in order to characterize the relation between the constants, $|\mathcal{X}|$, Γ_η, Γ_w, etc., which ensure convergence to some neighborhood of the maximum point satisfying $\|\nabla_\theta \eta(\theta)\|_2 = 0$. Theorem 4.7 states conditions on this convergence.

4.3.1 Relate the Algorithm to an ODE

In order to prove the convergence of this algorithm to the related ODE, we need to introduce the following assumption, which adds constraints to the iteration for w, and will be used in the sequel

Algorithm 1 TD AC Algorithm

Given:

- An MDP with a finite set $\mathcal{X}$ of states satisfying Assumption 3.2.
- An actor with a parametrized policy $\mu(u|x,\theta)$ satisfying Assumptions 3.4 and 4.1.
- A critic with a linear basis for $\tilde{h}(w)$, i.e., $\{\phi\}_{i=1}^{L}$, satisfying Assumption 4.3.
- A set H, a constant B_w, and an operator Ψ_w according to Definition 4.5.
- Step parameters Γ_η and Γ_h.
- Choose a TD parameter $0 \leq \lambda < 1$.

For step $n = 0$:

- Initiate the critic and the actor variables: $\tilde{\eta}_0 = 0$,$w_0 = 0$, $e_0 = 0$, $\theta_0 = 0$.

For each step $n = 1, 2, \ldots$

Critic: Calculate the estimated TD and eligibility trace

$$\begin{aligned} \tilde{\eta}_{n+1} &= \tilde{\eta}_n + \gamma_n \Gamma_\eta \left(r(x_n) - \tilde{\eta}_n \right) \\ \tilde{h}(x, w_n) &= w_n' \phi(x), \\ \tilde{d}\left(x_n, x_{n+1}, w_n\right) &= r(x_n) - \tilde{\eta}_n + \tilde{h}(x_{n+1}, w_n) - \tilde{h}(x_n, w_n), \\ e_n &= \lambda e_{n-1} + \phi\left(x_n\right). \end{aligned} \tag{23}$$

Set,

$$w_{n+1} = w_n + \gamma_n \Gamma_w \tilde{d} Cramer's \left(x_n, x_{n+1}, w_n\right) e_n \tag{24}$$

Actor:

$$\theta_{n+1} = \theta_n + \gamma_n \psi(x_n, u_n, \theta_n) \tilde{d}_n(x_n, x_{n+1}, w_n) \tag{25}$$

Project each component of w_{m+1} onto H (see Definition 4.5)

to prove Theorem 4.6. This assumption may seem restrictive at first but in practice it is not. The reason is that we usually assume the bounds of the constraints to be large enough so the iterates practically do not reach those bounds. For example, under Assumption 3.2 and additional mild assumptions, it is easy to show that $h(\theta)$ is uniformly bounded for all $\theta \in \mathbb{R}^K$. As a result, there exist a constant bounding $w^*(\theta)$ for all $\theta \in \mathbb{R}^K$. Choosing constraints larger than this constant will not influence the algorithm performance.

Definition 4.5 *Let us denote by $\{w_i\}_{i=1}^L$ the components of w, and choose a positive constant B_w. We define the set $H \subset \mathbb{R}^K \times \mathbb{R}^L$ to be*

$$H \triangleq \{(\theta, w) \,|-\infty < \theta_i < \infty, \quad 1 \le i \le K, \quad -B_w \le w_j \le B_w, \quad 1 \le j \le L\},$$

and let Ψ_w be an operator which projects w onto H, i.e., for each Cramer's$1 \le j \le L$, $\Psi_w w_j = \max(\min(w_j, B_w), -B_w)$.

The following theorem identifies the ODE system which corresponds to Algorithm 1. The detailed proof is given in Appendix C.

Theorem 4.6 *Define the following functions:*

$$\begin{aligned} G(\theta) &= \Phi' \Pi(\theta) \sum_{m=0}^{\infty} \lambda^m P(\theta)^m, \\ D^{(x,u,y)}(\theta) &= \pi(x) P(u|x,\theta) P(y|x,u) \psi(x,u,\theta), \quad x,y \in \mathcal{X}, \quad u \in \mathcal{U}. \\ A(\theta) &= \Phi' \Pi(\theta) (M(\theta) - I) \Phi, \\ M(\theta) &= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m P(\theta)^{m+1}, \\ b(\theta) &= \Phi' \Pi(\theta) \sum_{m=0}^{\infty} \lambda^m P(\theta)^m (r - \eta(\theta)). \end{aligned} \tag{26}$$

Then,

1. Algorithm 1 converges to the invariant set of the following set of ODEs

$$\begin{cases} \dot{\theta} = \nabla_\theta \eta(\theta) + \sum_{x,y \in \mathcal{X} \times \mathcal{X}} D^{(x,u,y)}(\theta) \left(d(x,y,\theta) - \tilde{d}(x,y,w) \right), \\ \dot{w} = \Psi_w \left[\Gamma_w \left(A(\theta) w + b(\theta) + G(\theta)(\eta(\theta) - \tilde{\eta}) \right) \right], \\ \dot{\tilde{\eta}} = \Gamma_\eta (\eta(\theta) - \tilde{\eta}), \end{cases} \tag{27}$$

   with probability 1.

   (a) The functions in (26) are continuous with respect to θ.

4.3.2 Investigating the ODE Asymptotic Behavior

Next, we quantify the asymptotic behavior of the system of ODEs in terms of the various algorithmic parameters. The proof of the theorem appears in Appendix D.

Theorem 4.7 *Consider the constants Γ_η and Γ_w as defined in Algorithm 1, and the function approximation bound ϵ_{app} as defined in* (17)*. Setting*

$$B_{\nabla\eta} \triangleq \frac{B_{\Delta td1}}{\Gamma_w} + \frac{B_{\Delta td2}}{\Gamma_\eta} + B_{\Delta td3} \epsilon_{app},$$

where $B_{\Delta td1}$, $B_{\Delta td2}$, $B_{\Delta td3}$ are a finite constants depending on the MDP and agent parameters. Then, the ODE system (27) *satisfies*

$$\liminf_{t\to\infty} \|\nabla_\theta \eta(\theta_t)\| \leq B_{\nabla\eta}. \tag{28}$$

Theorem 4.7 has a simple interpretation. Consider the trajectory $\eta(\theta_t)$ for large times, corresponding to the asymptotic behavior of η_n. The result implies that the trajectory visits a neighborhood of a local maximum infinitely often. Although it may leave the local vicinity of the maximum, it is guaranteed to return to it infinitely often. This occurs, since once it leaves the vicinity, the gradient of η points in a direction which has a positive projection on the gradient direction, thereby pushing the trajectory back to the vicinity of the maximum. It should be noted that in simulation (reported below) the trajectory usually remains within the vicinity of the local maximum, rarely leaving it. We also observe that by choosing appropriate values for Γ_η and Γ_w we can control the size of the ball to which the algorithm converges.

The key idea required to prove the Theorem is the following argument. If the trajectory does not satisfy $\|\nabla\eta(\theta)\|_2 \leq B_{\nabla\eta}$, we have $\dot{\eta}(\theta) > \epsilon$ for some positive ϵ. As a result, we have a monotone function which increases to infinity, thereby contradicting the boundedness of $\eta(\theta)$. Thus, $\eta(\theta)$ must visit the set which satisfies $\|\nabla\eta(\theta)\|_2 \leq B_{\nabla\eta}$ infinitely often.

5. A Comparison to other convergence results

In this section, we point out the main differences between Algorithm 1, the first algorithm proposed by Bhatnagar et al. (2008b) and the algorithms proposed by Konda and Tsitsiklis (2003). The main dimensions along which we compare the algorithms are the time scale, the type of the TD signal, and whether the algorithm is on line or off line.

The Time Scale and Type of Convergence

As was mentioned previously, the algorithms of Bhatnagar et al. (2008b) and Konda and Tsitsiklis (2003) need to operate in two time scales. More precisely, this refers to the following situation. Denote the time step of the critic's iteration by γ_n^c and the time step of the actor's iteration by γ_n^a, we have $\gamma_n^c = o(\gamma_n^a)$, i.e.,

$$\lim_{n\to\infty} \frac{\gamma_n^c}{\gamma_n^a} = 0.$$

The use of two time scales stems from the need of the critic to give an accurate estimate of the state values (as in the work of Bhatnagar et al. (2008b)) or the state-action values (as in the work of Konda and Tsitsiklis (2003)) before the actor uses them.

In the algorithm proposed here, a single time scale is used for the three iterates of Algorithm 1. We have $\gamma_n^a = \gamma_n$ for the actor iterate, $\gamma_n^{c,\eta} = \Gamma_\eta \gamma_n$ for the critic's η_n iterate, and $\gamma_n^{c,w} = \Gamma_w \gamma_n$ for the critic's w iterate. Thus,

$$\lim_{n\to\infty} \frac{\gamma_n^{c,\eta}}{\gamma_n^a} = \Gamma_\eta,$$
$$\lim_{n\to\infty} \frac{\gamma_n^{c,w}}{\gamma_n^a} = \Gamma_w.$$

Due to the single time scale, Algorithm 1 hasthe potential to converge faster than algorithms based on two time scales, since both the actor and the critic may operate on the fast time scale. The drawback of Algorithm 1 is the fact that convergence to the optimal value cannot be guaranteed, as was proved by Bhatnagar et al. (2008b) and by Konda and Tsitsiklis (2003). Instead, convergence to a neighborhood in $\mathbb{R}^K$ around the optimal value is guaranteed. In order to make the neighborhood smaller, we need to choose Γ_η and Γ_w appropriately, as is stated in Theorem 4.7.

The TD Signal, the Information Passed Between the Actor and the Critic, and the Critic's Basis

The algorithm presented in Bhatnagar et al. (2008b) is essentially a TD(0) algorithm, while the algorithm in Konda and Tsitsiklis (2003) is TD(1), Our algorithm is a TD(λ) for $0 \leq \lambda < 1$. A major difference between the approaches in Bhatnagar et al. (2008b) and the present work, as compared to Konda and Tsitsiklis (2003), is the information passed from the critic to the actor. In the former cases, the information passed is the TD signal, while in the latter case the Q-value is passed. Additionally, in Bhatnagar et al. (2008b) and in Algorithm 1 the critic's basis functions do not change through the simulation, while in Konda and Tsitsiklis (2003) the critic's basis functions are changed in each iteration according to the actor's parameter θ. Finally, we comment that Bhatnagar et al. (2008b) introduced an additional algorithm, based on the so-called natural gradient, which led to improved convergence speed. In this work we limit ourselves to algorithms based on the regular gradient, and defer the incorporation of the natural gradient to future work. As stated in Section 1, our motivation in this work was the derivation of a single time scale online AC algorithm with guaranteed convergence, which may be applicable in a biological context. The more complex natural gradient approach seems more restrictive in this setting.

6. Simulations

We report empirical results applying Algorithm 1 to a set of abstract randomly constructed MDPs which are termed Average Reward Non-stationary Environment Test-bench or in short GARNET (Archibald et al. (1995)). GARNET problems comprise a class of randomly constructed finite MDPs serving as a test-bench for control and RL algorithms optimizing the average reward per stage. A GARNET problem is characterized in our case by four parameters and is denoted by GARNET(X, U, B, σ). The parameter X is the number of states in the MDP, U is the number of actions, B is the branching factor of the MDP, i.e., the number of non-zero entries in each line of the MDP's transition matrices, and σ is the variance of each transition reward.

We describe how a GARNET problem is generated. When constructing such a problem, we generate for each state a reward, distributed normally with zero mean and unit variance. For each state-action the reward is distributed normally with the state's reward as mean and variance σ^2. The transition matrix for each action is composed of B non-zero terms in each line which sum to one.

We note that a comparison was carried out by Bhatnagar et al. (2008b) between their algorithm and the algorithm of Konda and Tsitsiklis (2003). We therefore compare our results directly to the more closely related former approach (see also Section 5).

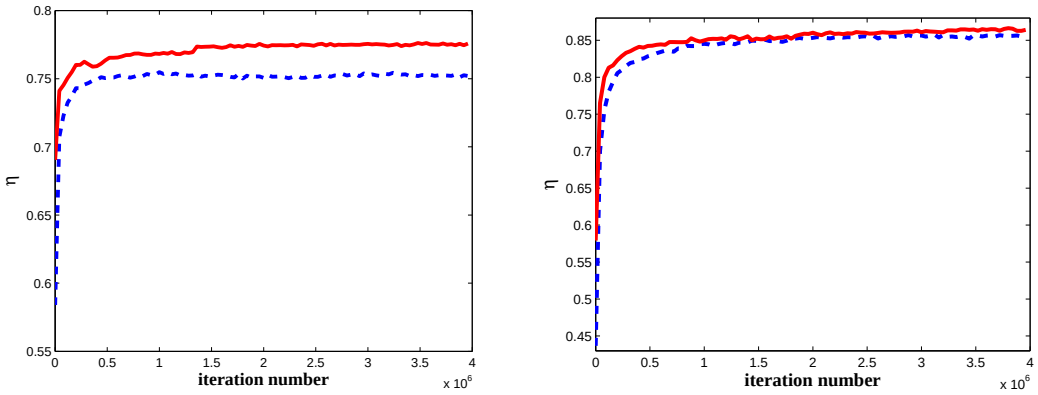
We consider the same GARNET problems as those simulated by Bhatnagar et al. (2008b). For completeness, we provide here the details of the simulation. For the critic's feature vector, we use a linear function approximation $\tilde{h}(x, w) = \phi(x)'w$, where $\phi(x) \in \{0, 1\}^L$, and define l to be the number nonzero values in $\phi(x)$. The nonzero values are chosen uniformly at random, where any two states have different feature vectors. The actor's feature vectors are of size $L \times |\mathcal{U}|$, and are constructed as

$$\xi(x, u) \triangleq (\overbrace{0, \ldots, 0}^{L\times(u-1)}, \phi(x), \overbrace{0, \ldots, 0}^{L\times(|\mathcal{U}|-u)}),$$

$$\mu(u|x, \theta) = \frac{e^{\theta'\xi(x,u)}}{\sum_{u'\in\mathcal{U}} e^{\theta'\xi(x,u')}}.$$

Bhatnagar et al. (2008b) reported simulation results for two GARNET problems: GARNET$(30, 4, 2, 0.1)$ and GARNET$(100, 10, 3, 0.1)$. For the GARNET$(30, 4, 2, 0.1)$ problem, Bhatnagar et al. (2008b) used critic steps $\gamma_n^{c,w}$ and $\gamma_n^{c,\eta}$, and actor steps γ_n^a, where

$$\gamma_n^{c,w} = \frac{100}{1000 + n^{2/3}}, \quad \gamma_n^{c,\eta} = 0.95\gamma_n^{c,w}, \quad \gamma_n^{a,\eta} = \frac{1000}{100000 + n},$$



(a) (b)

Figure 2: Simulation results applying Algorithm 1 (red solid line) and algorithm 1 of Bhatnagar et al. (2008b) (blue dashed line) on a GARNET$(30, 4, 2, 0.1)$ problem (a) and on GARNET$(100, 10, 3, 0.1)$ problem (b). Standard errors of the mean (suppressed for visibility) are of the order of 0.04.

and for GARNET$(100, 10, 3, 0.1)$ the steps were

$$\gamma_n^{c,w} = \frac{10^5}{10^6 + n^{2/3}}, \quad \gamma_n^{c,\eta} = 0.95\gamma_n^{c,w}, \quad \gamma_n^{a,\eta} = \frac{10^6}{10^8 + n}.$$

In our simulations we used a single time scale, γ_n, which was equal to $\gamma_n^{c,w}$ as used by Bhatnagar et al. (2008b). The basis parameters for GARNET$(30, 4, 2, 0.1)$ were $L = 8$ and $l = 3$, where for GARNET$(100, 10, 3, 0.1)$ they were $L = 20$ and $l = 5$.

In Figures 2 we show results of applying Algorithm 1 (solid line) and algorithm 1 from Bhatnagar et al. (2008b) (dashed line) on GARNET$(30, 4, 2, 0.1)$ and GARNET$(100, 10, 3, 0.1)$ problems. Each graph in Figure 2, represents an average of 100 independent simulations. Note that an agent with a uniform action selection policy will attain an average reward per stage of zero in these problems. Figure 3 presents similar results for GARNET$(30, 15, 15, 0.1)$. We see from these results that in all simulations, during the initial phase, Algorithm 1 converges faster than algorithm 1 from Bhatnagar et al. (2008b). The long term behavior is problem-dependent, as can be seen by comparing figures 2 and 3; specifically, in Figure 2 the present algorithm converges to a higher value than Bhatnagar et al. (2008b), while the situation is reversed in Figure 3. We refer the reader toMokkadem and Pelletier (2006) for careful discussion of convergence rates for two time scales algorithms; a corresponding analysis of convergence rates for single time scale algorithms is currently an open problem.

The results displayed here suggest a possible avenue for combining both algorithms. More concretely, using the present approach may lead to faster initial convergence due to the single time scale setting, which allows both the actor and the critic to evolve rapidly, while switching smoothly to a two time scales approach as in (Bhatnagar et al. (2008b)) will lead to asymptotic convergence to a point rather than to a region. This type of approach is reminiscent of the quasi-Newton algorithms in optimization, and is left for future work. As discussed in Section 5, we do not consider the natural gradient based algorithms from Bhatnagar et al. (2008b) in this comparative study.

7. Discussion and Future Work

We have introduced an algorithm where the information passed from the critic to the actor is the temporal difference signal, while the critic applies a TD(λ) procedure. A policy gradient approach

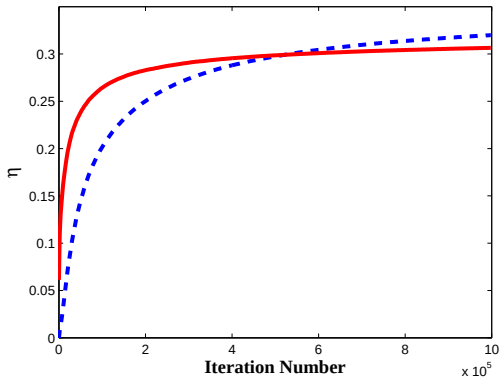


Figure 3: Simulation results applying Algorithm 1 (red solid line) and algorithm 1 of Bhatnagar et al. (2008b) (blue dashed line) on a GARNET$(30, 15, 15, 0.1)$ problem. Standard errors of the mean (suppressed for visibility) are of the order of 0.018.

was used in order to update the actor's parameters, based on a critic using linear function approximation. The main contribution of this work is a convergence proof in a situation where both the actor and the critic operate on the same time scale. The drawback of the extra flexibility in time scales is that convergence is only guaranteed to a neighborhood of a local maximum value of the average reward per stage. However, this neighborhood depends on parameters which may be controlled to improve convergence.

This work sets the stage for much future work. First, as observed above, the size of the convergence neighborhood is inversely proportional to the step sizes Γ_w and Γ_η. In other words, in order to reduce this neighborhood we need to select larger values of Γ_w and Γ_η. This on the other hand increases the variance of the algorithm. Therefore, further investigation of methods which reduce this variance are needed. However, the bounds used throughout are clearly rather loose, and cannot be effectively used in practical applications. Obviously, improving the bounds, and conducting careful numerical simulations in order to obtain a better practical understanding of the influence of the different algorithmic parameters, is called for. In addition, there is clearly room for combining the advantages of our approach with those of AC algorithms for which convergence to a single point is guaranteed, as discussed in Section 6,

From a biological point of view, our initial motivation to investigate TD based AC algorithms stemmed from questions related to the implementation of RL in the mammalian brain. Such a view is based on an interpretation of the transient activity of the neuromodulator dopamine as a TD signal (e.g., Schultz (2002)). Recent evidence suggested that the dorsal and ventral striatum may implement the actor and the critic, respectively (e.g., Daw et al. (2006)). We believe that theoretical models such as (Bhatnagar et al. (2008b)) and Algorithm 1 may provide, even if partially, a firm foundation to theories at the neural level. Some initial attempts in a neural setting (using direct policy gradient rather than AC based approaches) have been made by Baras and Meir (2007) and Florian (2007). Such an approach may lead to functional insights as to how an AC paradigm may be implemented at the cellular level of the basal ganglia and cortex. An initial demonstration was given by DiCastro et al. (2008).

From a theoretical perspective many issues remain open. First, strengthening Theorem (4.7) by replacing liminf by lim would clearly be useful. Second, extending the recent convergence rate results in Mokkadem and Pelletier (2006) to the single time scale case is an important challenging problem. Third, systematically combining the advantages of single time scale convergence (fast initial dynamics) and two time scales approaches (convergence to a point) would clearly be beneficial.

Acknowledgment The authors are grateful to Mohammad Ghavamzadeh for sending them a copy of Bhatnagar et al. (2008b) prior to publication, and to the anonymous reviewers for their helpful comments. The work of R. Meir was partially supported by an ISF Converging Technologies grant, and by ISF grant 665/08.

APPENDIX

Appendix A. Proofs of Results from Section 3

A.1 Proof of Lemma 3.6

1. Looking at (1) we see that $P(y|x,\theta)$ is a compound function of an integral and a twice differentiable function, $\mu(y|x,\theta)$, with bounded first and second derivatives according to Assumption 3.4. Therefore, $P(y|x,\theta)$ is a twice differentiable function with bounded first and second derivatives for all $\theta \in \mathbb{R}^K$.

2. According to Lemma 3.3, for each $\theta \in \mathbb{R}^K$ we have a unique solution to the following non-homogeneous linear equation system in $\{\pi(i|\theta)\}_{i=1}^{|\mathcal{X}|}$,

$$\begin{cases} \sum_{i=1}^{|\mathcal{X}|} \pi(i|\theta) P(j|i,\theta) = \pi(j|\theta), \quad j = 1,\ldots,|\mathcal{X}|-1, \\ \sum_{i=1}^{|\mathcal{X}|} \pi(i|\theta) = 1, \end{cases} \tag{29}$$

or in matrix form $M(\theta)\pi(\theta) = b$. By Assumption 3.2, the equation system (29) is invertible, therefore, $\det[M(\theta)] > 0$. This holds for all $P(\theta) \in \bar{P}$, thus, there exists a positive constant, b_M, which uniformly lower bounds $\det[M(\theta)]$ for all $\theta \in \mathbb{R}^K$.Thus, using Cramer's rule we have

$$\pi(i|\theta) = \frac{Q(i,\theta)}{\det[M(\theta)]},$$

where $Q(i,\theta)$ is a finite polynomial of $\{P(j|i,\theta)\}_{i,j\in\mathcal{X}}$ of at most degree $|\mathcal{X}|$ and with at most $|\mathcal{X}|!$ terms. Writing $\partial\pi(x|\theta)/\partial\theta_i$ explicitly gives

$$\begin{aligned} \left|\frac{\partial\pi(x|\theta)}{\partial\theta_i}\right| &= \left|\frac{\det[M(\theta)]\frac{\partial}{\partial\theta_i}Q(i,\theta) - Q(i,\theta)\frac{\partial}{\partial\theta_i}\det[M(\theta)]}{\det[M(\theta)]^2}\right| \\ &\leq \left|\frac{\frac{\partial}{\partial\theta_i}Q(i,\theta)}{\det[M(\theta)]}\right| + \left|\frac{Q(i,\theta)\frac{\partial}{\partial\theta_i}\det[M(\theta)]}{\det[M(\theta)]^2}\right| \\ &\leq \frac{|\mathcal{X}|\cdot|\mathcal{X}|!\cdot B_{P_1}}{b_M} + \frac{(|\mathcal{X}|\cdot|\mathcal{X}|!)\cdot B_{P_1}}{b_M^2}, \end{aligned}$$

which gives the desired bound. Following similar steps we can show the boundedness of the second derivatives.

3. The average reward per stage, $\eta(\theta)$ is a linear combination of $\{\pi(i|\theta)\}_{i=1}^{|\mathcal{X}|}$, with bounded coefficients by assumption 3.1. Therefore, using section 2, $\eta(\theta)$ is twice differentiable with bounded first and second derivatives for all $\theta \in \mathbb{R}^K$.

4. Since $\pi(x|\theta)$ is the stationary distribution of a recurrent MC, according to Assumption 3.2 there is a positive probability to be in each state $x \in \mathcal{X}$. This applies to the closure of $\mathcal{P}$. Thus, there exist a positive constant b_π such that $\pi(x|\theta) \geq b_\pi$.

A.2 Proof of Lemma 3.7

1. We recall the Poisson equation (5). We have the following system of linear equations in $\{h(x|\theta)\}_{x\in\mathcal{X}}$, namely,

$$\begin{cases} h(x|\theta) = r(x) - \eta(\theta) + \sum_{y\in\mathcal{X}} P(y|x,\theta)h(y|\theta), \quad \forall x \in \mathcal{X}, x \neq x^*, \\ h(x^*|\theta) = 0. \end{cases} \tag{30}$$

   or in matrix form $N(\theta)h(\theta) = c$. Adding the equation $h(x^*|\theta) = 0$ yields a unique solution for the system (Bertsekas (2006), Vol. 1, Prop. 7.4.1). Thus, using Cramer's rule we have $h(x|\theta) = R(x,\theta)/\det[N(\theta)]$, where $R(x,\theta)$ and $\det[N(\theta)]$ are polynomial function of entries in $N(\theta)$, which are bounded and have bounded first and second derivatives according to Lemma 3.6. Continuing in the same steps of Lemma 3.6 proof, we conclude that $h(x|\theta)$ and its two first derivatives for all $x \in \mathcal{X}$ and for all $\theta \in \mathbb{R}^K$.

2. Trivially, by (6) and the previous section the result follows.

Appendix B. Proof of Theorem 4.2

We begin with a Lemma which was proved in (Marbach and Tsitsiklis (1998)). It relates the gradient of the average reward per stage to the differential value function.

Lemma B.1 *The gradient of the average reward per stage can be expressed by*

$$\nabla_\theta \eta(\theta) = \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)h(y,\theta). \tag{31}$$

For completeness, we present a proof,which will be used in the sequel.
Proof We begin with Poisson's equation (5) in vector form

$$h(\theta) = \bar{r} - e\eta(\theta) + P(\theta)h(\theta),$$

where e is a column vector of 1's. Taking the derivative with respect to θ and rearranging yields

$$e\nabla_\theta \eta(\theta) = -\nabla_\theta h(\theta) + \nabla_\theta P(\theta)h(\theta) + P(\theta)\nabla_\theta h(\theta).$$

Multiplying the left hand side of the last equation by the stationary distribution $\pi(\theta)'$ yields

$$\begin{aligned} \nabla_\theta \eta(\theta) &= -\pi(\theta)'\nabla_\theta h(\theta) + \pi(\theta)'\nabla_\theta P(\theta)h(\theta) + \pi(\theta)' P(\theta)\nabla_\theta h(\theta) \\ &= -\pi(\theta)'\nabla_\theta h(\theta) + \pi(\theta)'\nabla_\theta P(\theta)h(\theta) + \pi(\theta)'\nabla_\theta h(\theta) \\ &= \pi(\theta)'\nabla_\theta P(\theta)h(\theta). \end{aligned}$$

Expressing the result explicitly we obtain

$$\begin{aligned}
\nabla_\theta \eta(\theta) &= \sum_{x,y\in\mathcal{X}} P(x)\nabla_\theta P(y|x,\theta)h(y,\theta) \\
&= \sum_{x,y\in\mathcal{X}} P(x)\nabla_\theta \left(\sum_u \left(P(y|x,u)\mu(u|x,\theta)\right)\right) h(y,\theta) \\
&= \sum_{x,y\in\mathcal{X}} P(x)\sum_u \left(P(y|x,u)\nabla_\theta\mu(u|x,\theta)\right) h(y,\theta) \\
&= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(y|x,u)P(x)\nabla_\theta\mu(u|x,\theta)h(y,\theta) \\
&= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(y|x,u)\mu(u|x,\theta)P(x)\frac{\nabla_\theta\mu(u|x,\theta)}{\mu(u|x,\theta)}h(y,\theta) \\
&= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)h(y,\theta).
\end{aligned} \tag{32}$$

■

Based on this, we can now prove Theorem 4.2. We start with the result in (32).

$$\begin{aligned}
\nabla_\theta \eta(\theta) &= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)h(y,\theta). \\
&= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)\left(h(y,\theta) - h(x,\theta) + \bar{r}(x) - \eta(\theta) + f(x)\right) \\
&\quad - \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)\left(-h(x,\theta) + \bar{r}(x) - \eta(\theta) + f(x)\right) \\
&= \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)\left(d(x,y,\theta) + f(x)\right) \\
&\quad - \sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)\left(-h(x,\theta) + \bar{r}(x) - \eta(\theta) + f(x)\right)
\end{aligned}$$

In order to complete the proof, we show that the second term equals 0. We define $F(x,\theta) \triangleq -h(x|\theta) + \bar{r}(x) - \eta(\theta) + f(x)$ and obtain

$$\begin{aligned}
\sum_{x,y\in\mathcal{X},u\in\mathcal{U}} P(x,u,y,\theta)\psi(x,u,\theta)F(x,\theta) &= \sum_{x\in\mathcal{X}} \pi(x,\theta)F(x,\theta) \sum_{u\in\mathcal{U},y\in\mathcal{X}} \nabla_\theta P(y|x,u,\theta) \\
&= 0.
\end{aligned}$$

Appendix C. Proof of Theorem 4.6

As mentioned earlier, we use Theorem 6.1.1 of Kushner and Yin (1997). We start by describing the setup of the theorem and the main result. Then, we show that the required assumptions hold in our case.

C.1 Setup, Assumptions and Theorem 6.1.1 of Kushner and Yin (1997).

In this section we describe briefly but accurately the conditions for Theorem 6.1.1 of Kushner and Yin (1997) and state the main result. We consider the following stochastic iteration

$$y_{n+1} = \Pi_H[y_n + \gamma_n Y_n], \tag{33}$$

where Y_n is a vector of "observations" at time n, and Π_H is a constraint operator as defined in Definition 4.5. Recall that $\{x_n\}$ is a Markov chain. Based on this, define $\mathcal{F}_n$ to be the σ-algebra

$$\begin{aligned}\mathcal{F}_n &\triangleq \sigma\{y_0, Y_{i-1}, x_i \,|i \leq n\} \\ &= \sigma\{y_0, Y_{i-1}, x_i, y_i \,|i \leq n\},\end{aligned}$$

and

$$\bar{\mathcal{F}}_n \triangleq \sigma\{y_0, Y_{i-1}, y_i \,|i \leq n\}.$$

The difference between the σ-algebras is the sequence $\{x_n\}$. Define the conditioned average iterate

$$g_n(y_n, x_n) \triangleq \mathrm{E}[Y_n \,|\mathcal{F}_n],$$

and the corresponding *martingale difference noise*

$$\delta M_n \triangleq Y_n - \mathrm{E}[Y_n \,|\mathcal{F}_n].$$

Thus, we can write the iteration as

$$y_{n+1} = y_n + \gamma_n (g_n(y_n, x_n) + \delta M_n + Z_n),$$

where Z_n is a reflection term which forces the iterate to the nearest point in the set H whenever the iterates leaves it (see Kushner and Yin (1997) for details). Next, set

$$\bar{g}(y) \triangleq \mathrm{E}\left[g_n(y, x_n) \,\middle|\bar{\mathcal{F}}_n\right].$$

Later, we will see that the sum of the sequence $\{\delta M_n\}$ converges to 0, and the r.h.s of the iteration behaves approximately as a the function $\bar{g}(y)$, which yields the corresponding ODE, i.e.,

$$\dot{y} = \bar{g}(y).$$

The following ODE method will show that the asymptotic behavior of the iteration is equal to the asymptotic behavior of the corresponding ODE.

Define the auxiliary variable

$$t_n \triangleq \sum_{k=0}^{n-1} \gamma_k,$$

and the monotone piecewise constant auxiliary function

$$m(t) = \{n \,|t_n \leq t < t_{n+1}\}.$$

The following assumption, taken from Section 6.1 of Kushner and Yin (1997), is required to establish the basic Theorem. An interpretation of the assumption follows its statement.

Assumption C.1 *Assume that*

1. The coefficients $\{\gamma_n\}$ satisfy $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\lim_{n\to\infty} \gamma_n = 0$.
   (a) $\sup_n \mathrm{E}[\|Y_n\|] < \infty$.
   (b) $g_n(y_n, x)$ is continuous in y_n for each x and n.
   (c) For each $\mu > 0$ and for some $T > 0$ there is a continuous function $\bar{g}(\cdot)$ such that for each y

$$\lim_{n\to\infty} \Pr\left(\sup_{j\geq n} \max_{0\leq t\leq T} \left\| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i (g_n(y, x_i) - \bar{g}(y)) \right\| \geq \mu\right) = 0.$$

(d) For each $\mu > 0$ and for some $T > 0$ we have

$$\lim_{n\to\infty} \Pr\left(\sup_{j\geq n} \max_{0\leq t\leq T} \left\| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i \delta M_i \right\| \geq \mu\right) = 0.$$

(e) There are measurable and non-negative functions $\rho_3(y)$ and $\rho_{n4}(x)$such that

$$\|g_n(y_n, x)\| \leq \rho_3(y)\,\rho_{n4}(x)$$

where $\rho_3(y)$ is bounded on each bounded y-set , and for each $\mu > 0$ we have

$$\lim_{\tau\to 0} \lim_{n\to\infty} \Pr\left(\sup_{j\geq n} \sum_{i=m(j\tau)}^{m(j\tau+\tau)-1} \gamma_i \rho_{n4}(x_i) \geq \mu\right) = 0.$$

(f) There are measurable and non-negative functions $\rho_1(y)$ and $\rho_{n2}(x)$such that $\rho_1(y)$ is bounded on each bounded y-set and

$$\|g_n(y_1, x) - g_n(y_2, x)\| \leq \rho_1(y_1 - y_2)\,\rho_{n2}(x),$$

where

$$\lim_{y\to 0} \rho_1(y) = 0,$$

and

$$\Pr\left(\limsup_j \sum_{i=j}^{m(t_j+\tau)} \gamma_i \rho_{i2}(x_i) < \infty\right) = 1.$$

The conditions of Assumption C.1 are quite general but can be interpreted as follows. Assumptions C.1.1-3 are straightforward. Assumption C.1.4 is reminiscent of ergodicity, which is used to replace the state-dependent function $g_n(\cdot,\cdot)$ with the state-independent of state function $\bar{g}(\cdot)$, whereas Assumption C.1.5 states that the martingale difference noise converges to 0 in probability. Assumptions C.1.6 and C.1.7 ensure that the function $g_n(\cdot,\cdot)$ is not unbounded and satisfies a Lipschitz condition.

The following Theorem, adapted from Kushner and Yin (1997), provides the main convergence result required. The remainder of this appendix shows that the required conditions in Assumption C.1 hold.

Theorem C.2 *(Adapted from Theorem 6.1.1 in Kushner and Yin (1997)) Assume that algorithm 1, and Assumption C.1 hold. Then y_n converges to some invariant set of the projected ODE*

$$\dot{y} = \Pi_H[\bar{g}(y)].$$

Thus, the remainder of this section is devoted to showing that Assumptions C.1.1-C.1.7 are satisfied.

For future purposes, we express Algorithm 1 using the augmented parameter vector y_n

$$y_n \triangleq (\theta_n' \quad w_n' \quad \tilde{\eta}_n')', \quad \theta_n \in \mathbb{R}^K, \quad w_n \in \mathbb{R}^L, \quad \tilde{\eta}_n \in \mathbb{R}. \tag{34}$$

The components of Y_n are determined according to (27). The corresponding sub-vectors of $\bar{g}(y_n)$ will be denoted by

$$\bar{g}(y_n) = \left[\bar{g}(\theta_n)' \quad \bar{g}(w_n)' \quad \bar{g}(\tilde{\eta}_n)'\right]' \in \mathbb{R}^{K+L+1},$$

and similarly

$$g_n(y_n, x_n) = \left[g_n(\theta_n, x_n)' \quad g_n(w_n, x_n)' \quad g_n(\tilde{\eta}_n, x_n)'\right]' \in \mathbb{R}^{K+L+1}.$$

We begin by examining the components of $g_n(y_n, x_n)$ and $\bar{g}(y_n)$. The iterate $g_n(\tilde{\eta}_n, x_n)$ is

$$\begin{aligned} g_n(\tilde{\eta}_n, x_n) &= \mathrm{E}\left[\Gamma_\eta\left(r(x_n) - \tilde{\eta}_n\right)\middle|\mathcal{F}_n\right] \\ &= \Gamma_\eta\left(r(x_n) - \tilde{\eta}_n\right), \end{aligned} \tag{35}$$

and since there is no dependence on x_n we have also

$$\bar{g}(\tilde{\eta}_n) = \Gamma_\eta\left(\eta(\theta) - \tilde{\eta}_n\right).$$

The iterate $g_n(w_n, x_n)$ is

$$\begin{aligned} g_n(w_n, x_n) &= \mathrm{E}\left[\Gamma_w \tilde{d}(x_n, x_{n+1}, w_n)\, e_n \middle| \mathcal{F}_n\right] \\ &= \mathrm{E}\left[\Gamma_w \sum_{k=0}^{\infty} \lambda^k \phi(x_{n-k})\left(r(x_n) - \tilde{\eta}_n + \phi(x_{n+1})' w_n - \phi(x_n)' w_n\right)\middle| \mathcal{F}_n\right] \\ &= \Gamma_w \sum_{k=0}^{\infty} \lambda^k \phi(x_{n-k})\left(r(x_n) - \tilde{\eta}_n + \sum_{y\in\mathcal{X}} P(y|x_n, \theta_n)\,\phi(y)' w_n - \phi(x_n)' w_n\right), \end{aligned} \tag{36}$$

and the iterate $\bar{g}(w_n)$ is

$$\begin{aligned} \bar{g}(w_n) &= \mathrm{E}\left[g_n(w_n, x_n)\middle|\bar{\mathcal{F}}_n\right] \\ &= \mathrm{E}\left[\Gamma_w \sum_{k=0}^{\infty} \lambda^k \phi(x_{n-k})\left(r(x_n) - \tilde{\eta}_n + \sum_{y\in\mathcal{X}} P(y|x_n, \theta_n)\,\phi(y)' w_n - \phi(x_n)' w_n\right)\middle|\bar{\mathcal{F}}\right] \\ &= \Gamma_w \sum_{k=0}^{\infty} \lambda^k \sum_{x\in\mathcal{X}} \pi(x)\,\phi(x) \sum_{z\in\mathcal{X}} \left[P^k\right]_{xz}\left(r(z) - \tilde{\eta}_n + \sum_{y\in\mathcal{X}} P(y|z, \theta_n)\,\phi(z)' w_n - \phi(y)' w_n\right), \end{aligned}$$

which, following, Bertsekas and Tsitsiklis (1996) section 6.3, can be written in matrix form

$$\bar{g}(w_n) = \Phi'\Pi(\theta_n)\left((1-\lambda)\sum_{k=0}^{\infty}\lambda^k P^{k+1} - I\right)\Phi w_n + \Phi'\Pi(\theta_n)\sum_{k=0}^{\infty}\lambda^k P^k\left(r - \tilde{\eta}_n\right).$$

With some further algebra we can express this using (26),

$$\bar{g}(w_n) = A(\theta_n)\, w_n + b(\theta_n) + G(\theta_n)\left(\eta(\theta_n) - \tilde{\eta}_n\right).$$

Finally, the iterate $g_n(\theta_n, x_n)$ is

$$\begin{aligned} g_n(\theta_n, x_n) &= \mathrm{E}\left[\tilde{d}(x_n, x_{n+1}, w_n)\,\psi(x_n, u_n, \theta_n)\middle|\mathcal{F}_n\right] \\ &= \mathrm{E}\left[d(x_n, x_{n+1}, \theta_n)\,\psi(x_n, u_n, \theta_n)\middle|\mathcal{F}_n\right] \\ &\quad + \mathrm{E}\left[\left(\tilde{d}(x_n, x_{n+1}, w_n) - d(x_n, x_{n+1}, \theta_n)\right)\psi(x_n, u_n, \theta_n)\middle|\mathcal{F}_n\right] \\ &= \mathrm{E}\left[d(x_n, x_{n+1}, \theta_n)\,\psi(x_n, u_n, \theta_n)\middle|\mathcal{F}_n\right] \\ &\quad + \sum_{z\in\mathcal{X}} P(z|x_n)\,\psi(x_n, u_n, \theta_n)\left(\tilde{d}(x_n, z, w_n) - d(x_n, z, \theta_n)\right), \end{aligned} \tag{37}$$

and

$$\begin{aligned} \bar{g}(\theta_n) &= \mathrm{E}\left[\tilde{d}(x_n, x_{n+1}, w_n)\,\psi(x_n, u_n, \theta_n)\middle|\bar{\mathcal{F}}_n\right] \\ &= \mathrm{E}\left[d(x_n, x_{n+1}, \theta_n)\,\psi(x_n, u_n, \theta_n)\middle|\bar{\mathcal{F}}_n\right] + \mathrm{E}\left[\left(\tilde{d}(x_n, x_{n+1}, w_n) - d(x_n, x_{n+1}, \theta_n)\right)\psi(x_n, u_n, \theta_n)\middle|\bar{\mathcal{F}}_n\right] \\ &= \nabla\eta(\theta_n) + \sum_{x,y\in\mathcal{X}}\sum_{u\in\mathcal{U}} \pi(x)\,P(u|x, \theta_n)\,P(y|x, u)\,\psi(x, u, \theta_n)\left(\tilde{d}(x, y, w_n) - d(x, y, \theta_n)\right). \end{aligned}$$

Next, we show that the required assumptions hold.

C.2 Satisfying Assumption C.1.2

We need to show that $\sup_n \mathrm{E}\left[\|Y_n\|_2\right] < \infty$. Since later we need to show that $\sup_n \mathrm{E}\left[\|Y_n\|_2^2\right] < \infty$, and the proof of the second moment is similar to the proof of the first moment, we consider both moments here.

Lemma C.3 *The sequence $\tilde{\eta}_n$ is bounded w.p. 1, $\sup_n E\left[\|Y_n(\tilde{\eta}_n)\|_2\right] < \infty$, and $\sup_n E\left[\|Y_n(\tilde{\eta}_n)\|_2^2\right] < \infty$*

Proof We can choose M such that $\gamma_n \Gamma_\eta < 1$ for all $n > M$. Using Assumption 3.4 for the boundedness of the rewards, we have

$$\begin{aligned}
\tilde{\eta}_{n+1} &= (1-\gamma_n\Gamma_\eta)\tilde{\eta}_m + \gamma_n\Gamma_\eta r(x_n) \\
&\leq (1-\gamma_n\Gamma_\eta)\tilde{\eta}_n + \gamma_n\Gamma_\eta B_r \\
&\leq \begin{cases} \tilde{\eta}_n & \text{if } \tilde{\eta}_n > B_r, \\ B_r & \text{if } \tilde{\eta}_n \leq B_r, \end{cases} \\
&\leq \max\{\tilde{\eta}_n, B_r\},
\end{aligned} \tag{38}$$

which means that each iterate is bounded above by the previous iterate or by a constant. We denote this bound by $B_{\tilde{\eta}}$. Using similar arguments we can prove that $\tilde{\eta}_n$ is bounded below, and the first part of the lemma is proved. Since $\tilde{\eta}_{n+1}$ is bounded the second part follows trivially. ■

Lemma C.4 *We have $\sup_n E\left[\|Y_n(w_n)\|_2^2\right] < \infty$ and $\sup_n E\left[\|Y_n(w_n)\|_2\right] < \infty$*

Proof For the first part we have

$$\begin{aligned}
\mathrm{E}\left[\|Y_n(w_n)\|_2^2\right] &= \mathrm{E}\left[\left\|\Gamma_w \tilde{d}(x_n, x_{n+1}, w_n)\, e_n\right\|_2^2\right] \\
&= \Gamma_w^2 \mathrm{E}\left[\left\|\sum_{k=0}^{\infty} \lambda^k \phi(x_{n-k})\left(r(x_n) - \tilde{\eta}_n + \phi(x_{n+1})' w_n - \phi(x_n)' w_n\right)\right\|_2^2\right] \\
&\overset{(a)}{\leq} \Gamma_w^2 \mathrm{E}\left[\sum_{k=0}^{\infty} \lambda^k \left\|\phi(x_{n-k})\left(r(x_n) - \tilde{\eta}_n + \phi(x_{n+1})' w_n - \phi(x_n)' w_n\right)\right\|_2\right]^2 \\
&\leq \Gamma_w^2 \mathrm{E}\left[\sup_k \left\|\phi(x_{n-k})\left(r(x_n) - \tilde{\eta}_n + \phi(x_{n+1})' w_n - \phi(x_n)' w_n\right)\right\|_2 \sum_{k=0}^{\infty} \lambda^k\right]^2 \\
&\overset{(b)}{\leq} \frac{4\Gamma_w^2}{(1-\lambda)^2} \|\phi(x_{n-k})\|_2^2 \left(|r(x_n)|^2 + |\tilde{\eta}_n|^2 + \|\phi(x_{n+1})\|_2^2 \cdot \|w_n\|_2^2 + \|\phi(x_n)\|_2^2 \cdot \|w_n\|_2^2\right) \\
&\leq \frac{4\Gamma_w^2}{(1-\lambda)^2} B_\phi^2 \left(B_r^2 + B_{\tilde{\eta}}^2 + 2B_\phi^2 B_w^2\right),
\end{aligned}$$

where we used the triangle inequality in (a) and the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ in (b). The bound $\sup_n \mathrm{E}\left[\|Y_n(w_n)\|_2\right] < \infty$ follows directly from the Cauchy-Schwartz inequality. ■

Lemma C.5 *We have $\sup_n E\left[\|Y_n(\theta_n)\|_2^2\right] < \infty$ and $\sup_n E\left[\|Y_n(\theta_n)\|_2\right] < \infty$. The proof proceeds as in Lemma C.4.*

Based on Lemmas C.3, C.4, and C.5 we can assert Assumption C.1.2

C.3 Satisfying Assumption C.1.3

Assumption C.1.3 requires the continuity of $g_n(y_n, x_n)$ for each n and x_n. Again, we show that this assumption holds for the three parts of the vector y_n.

Lemma C.6 *The function $g_n(\tilde{\eta}_n, x_n)$ is a continuous function of $\tilde{\eta}_n$ for each n and x_n.*

Proof Since $g_n(\tilde{\eta}_n, x_n) = \Gamma_\eta(r(x_n) - \tilde{\eta}_n)$ the claim follows. ■

Lemma C.7 *The function $g_n(w_n, x_n)$ is a continuous function of $\tilde{\eta}_n$, w_n, and θ_n for each n and x_n.*

Proof The function is

$$g_n(w_n, x_n) = \Gamma_w \sum_k^\infty \lambda^k \phi(x_{n-k}) \left(r(x_n) - \tilde{\eta}_n + \sum_{y \in \mathcal{X}} P(y|x_n, \theta_n) \phi(y)' w_n - \phi(x_n)' w_n \right).$$

The probability transition $\sum_{y \in \mathcal{X}} P(y|x_n, \theta_n)$ can be written as $\sum_{y \in \mathcal{X}, u \in \mathcal{U}} P(y|x_n, u_n) \mu(u_n|x_n, \theta_n)$. The function $\mu(u_n|x_n, \theta_n)$ is continuous in θ_n by Assumption 3.6, and thus $g_n(w_n, x_n)$ is continuous in $\tilde{\eta}_n$ and θ_n and the lemma follows. ■

Lemma C.8 *The function $g_n(\theta_n, x_n)$ is a continuous function of $\tilde{\eta}_n$, w_n, and θ_n for each n and x_n.*

Proof By definition, the function $g_n(\theta_n, x_n)$ is

$$\begin{aligned} g_n(\theta_n, x_n) &= \mathrm{E}\left[\tilde{d}(x_n, x_{n+1}, w_n) \psi(x_n, u_n, \theta_n) \middle| \mathcal{F}_n \right] \\ &= \frac{\nabla_\theta \mu(u_n|x_n, \theta_n)}{\mu(u_n|x_n, \theta_n)} \left(r(x_n) - \tilde{\eta}_n + \sum_{y \in \mathcal{X}} P(y|x_n, \theta_n) \phi(y)' w_n - \phi(x_n)' w_n \right) \end{aligned}$$

Using similar arguments to Lemma C.7 the claim holds. ■

C.4 Satisfying Assumption C.1.4

In this section we prove the following convergence result: for each $\mu > 0$ and for some $T > 0$ there is a continuous function $\bar{g}(\cdot)$ such that for each y

$$\lim_{n \to \infty} \Pr\left(\sup_{j \geq n} \max_{0 \leq t \leq T} \left\| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i \left(g_n(y, x_i) - \bar{g}(y) \right) \right\| \geq \mu \right). \tag{39}$$

We start by showing that there exist independent cycles of the algorithm since the underlying Markov chain is recurrent and aperiodic. Then, we show that the cycles behave as a martingale, thus Doob's inequality can be used. Finally we show that the sum in (39) converges to 0 w.p. 1. We start investigating the regenerative nature of the process.

Based on Lemma 3.2, there exists a recurrent state common to all $MC(\theta)$, denoted by x^*. We define the series of *hitting times* of the recurrent state x^* by $t_0 = 0, t_1, t_2, ...$, where t_m it the m-th time the agent hits the state x^*. Mathematically, we can define this series recursively by

$$t_{m+1} = \inf\{n | x_n = x^*, n > t_m\}, \quad t_0 = 0,$$

and $T_m \triangleq t_{m+1} - t_m$. Define the m-th cycle of the algorithm to be the set of times

$$\mathcal{T}_m \triangleq \{n | t_{m-1} \leq n < t_m\}, \tag{40}$$

and the corresponding trajectories

$$\mathcal{C}_m \triangleq \{x_n | n \in \mathcal{T}_m\}. \tag{41}$$

Define a function, $\varrho(k)$, which returns the cycle to which the time k belongs to, i.e.,

$$\varrho(k) \triangleq \{m | k \in \mathcal{T}_m\}.$$

We notice that based on Lemma 3.3, and using the *Regenerative Cycle Theorem* (see Brémaud (1999), pp. 87), the cycles $\mathcal{C}_m$ are independent of each other.

Next, we examine (39), and start by defining the following events:

$$\begin{aligned}
b_n^{(1)} &\triangleq \left\{\omega \left| \sup_{j\geq n} \max_{0\leq t\leq T} \left\| \sum_{i=m(jT)}^{m(jT+t)-1} \gamma_i \left(g_i(y,x_i) - \bar{g}(y)\right)\right\| \geq \mu \right.\right\}, \\
b_n^{(2)} &\triangleq \left\{\omega \left| \sup_{j\geq n} \sup_{k\geq m(jT)} \left\| \sum_{i=m(jT)}^{k} \gamma_i \left(g_i(y,x_i) - \bar{g}(y)\right)\right\| \geq \mu \right.\right\}, \\
b_n^{(3)} &\triangleq \left\{\omega \left| \sup_{j\geq n} \left\| \sum_{i=n}^{\infty} \gamma_i \left(g_i(y,x_i) - \bar{g}(y)\right)\right\| \geq \mu \right.\right\}.
\end{aligned}$$

It is easy to show that for each n we have $b_n^{(1)} \subset b_n^{(2)}$, thus,

$$\Pr\left(b_n^{(1)}\right) \leq \Pr\left(b_n^{(2)}\right). \tag{42}$$

It is easy to verify that the series $\left\{b_n^{(2)}\right\}$ is a subsequence of $\left\{b_n^{(3)}\right\}$. Thus, if we prove that $\lim_{n\to\infty} \Pr\left(b_n^{(3)}\right) = 0$, then $\lim_{n\to\infty} \Pr(b_n) = 0$, and using (42), Assumption C.1.4 holds.

Next, we examine the sum defining the event $b_n^{(3)}$, by splitting it a sum over cycles and a sum within each cycle. We can write it as following

$$\sum_{i=n}^{\infty} \gamma_i \left(g_i(y,x_i) - \bar{g}(y)\right) = \sum_{m=\varrho(n)}^{\infty} \sum_{i\in\mathcal{T}_m} \gamma_i \left(g_i(y,x_i) - \bar{g}(y)\right).$$

Denote $c_m \triangleq \sum_{j\in\mathcal{T}_m} \gamma_i \left(g_n(y,x_i) - \bar{g}(y)\right)$. Therefore, by the *Regenerative Cycle Theorem* (Brémaud (1999), pp. 87), c_m are independent random variables. Also,

$$\mathrm{E}[c_m] = \mathrm{E}\left[\sum_{i\in\mathcal{T}_m} \gamma_i \left(g_i(y,x_i) - \bar{g}(y)\right)\right] = \mathrm{E}\left[\mathrm{E}\left[\sum_{j\in\mathcal{T}_m} \gamma_i \left(g_n(y,x_i) - \bar{g}(y)\right) \middle| \mathcal{T}_m\right]\right] = 0.$$

We argue that c_m is square integrable. To prove this we need to show that the second moments of T_m and $\left(g_n(y,x_i) - \bar{g}(y)\right)$ are finite.

Lemma C.9

1. The first two moments of the random times $\{T_m\}$ are bounded above by a constant B_T, for all $\theta \in \mathbb{R}^K$ and for all m, $1 \leq m < \infty$.

(a) $\mathrm{E}\left[\left(g_n\left(y,x_i\right)-\bar{g}\left(y\right)\right)^2\right]\leq B_g$

(b) Define $\bar{\gamma}_m \triangleq \sup_{i\in\mathcal{T}_m}\gamma_i$, then $\sum_{m=0}^{\infty}\bar{\gamma}_m^2<\infty$.

(c) $\mathrm{E}\left[c_m^2\right]\leq\left(B_T B_g\right)^2$.

Proof ■

1. According to Assumption 3.2 and Lemma 3.3, each Markov chain in $\bar{\mathcal{P}}$ is recurrent. Thus, for each $\theta\in\mathbb{R}^K$ there exists a constant $\tilde{B}_T(\theta)$, $0<\tilde{B}_T(\theta)<1$, where for $k\leq|\mathcal{X}|$ we have

$$P(T_m=k|\theta_m)\leq\left(\tilde{B}_T(\theta_m)\right)^{\lfloor k/|\mathcal{X}|\rfloor},\quad 1\leq m<\infty,\quad 1\leq k<\infty, \tag{43}$$

where $\lfloor a\rfloor$ is the largest integer which is not greater than a. Otherwise, if for $k>|\mathcal{X}|$ we have $\tilde{B}_T(\theta_m)=1$ then the chain transitions equal 1 which contradicts the aperiodicity of the chains. Therefore,

$$\mathrm{E}\left[T_m|\,\theta_m\right]=\sum_{k=1}^{\infty}kP(T_m=k|\theta_m)\leq\sum_{k=1}^{\infty}k\left(\tilde{B}_T(\theta_m)\right)^{\lfloor k/|\mathcal{X}|\rfloor}=B_{T_1}(\theta_m)<\infty,$$

and

$$\mathrm{E}\left[T_m^2\big|\,\theta_m\right]=\sum_{k=1}^{\infty}k^2P(T_m=k|\theta_m)\leq\sum_{k=1}^{\infty}k^2\left(\tilde{B}_T(\theta_m)\right)^{\lfloor k/|\mathcal{X}|\rfloor}=B_{T_2}(\theta_m)<\infty.$$

Since the set $\bar{\mathcal{P}}$ is closed, by Assumption 3.2 the above holds for the closure of $\bar{\mathcal{P}}$ as well. Thus, there exists a constant B_T satisfying $B_T=\max\{\sup_\theta B_{T_1}(\theta),\sup_\theta B_{T_2}(\theta)\}<\infty$.

(a) The proof proceeds along the same lines as the proofs of lemmas C.3, C.4, and C.5.

(b) The result follows trivially since the sequence $\{\bar{\gamma}_m\}$ is subsequence of the summable sequence$\{\gamma_m\}$.

(c) By definition, for large enough m we have $\gamma_m\leq 1$. Therefore, we have

$$\begin{aligned}\mathrm{E}\left[c_m^2\right] &= \mathrm{E}\left[\left(\sum_{j\in\mathcal{T}_m}\gamma_j\left(g_n\left(y,x_j\right)-\bar{g}\left(y\right)\right)\right)^2\right]\\ &\leq \mathrm{E}\left[|\mathcal{T}_m|^2\left(\sup_j\gamma_j\right)^2\left(\sup_j\left(g_n\left(y,x_j\right)-\bar{g}\left(y\right)\right)\right)^2\right]\\ &\leq B_T^2B_g^2.\end{aligned}$$

Next, we conclude by showing that Assumption C.1.4 is satisfied. Define the process $d_n\triangleq\sum_{m=0}^{n}c_m$. This process is a martingale since the sequence $\{c_m\}$ is square integrable (by Lemma C.9) and satisfies

$\mathrm{E}\left[d_{m+1}|d_m\right] = d_m$. Using Doob's martingale inequality[1] we have

$$\begin{aligned}
\Pr\left(\sup_{k\geq n}\sum_{m=\varrho(n)}^{\varrho(k)}\sum_{j\in\mathcal{T}_m}\gamma_i\left(g_n\left(y,x_i\right)-\bar{g}\left(y\right)\right)\geq\mu\right) &\leq \lim_{n\to\infty}\frac{\mathrm{E}\left[\left(\sum_{m=\varrho(n)}^{\infty}\sum_{j\in\mathcal{T}_m}\gamma_j\left(g_n\left(y,x_j\right)-\bar{g}\left(y\right)\right)\right)^2\right]}{\mu^2} \\
&= \lim_{n\to\infty}\frac{\sum_{m=\varrho(n)}^{\infty}\mathrm{E}\left[\left(\sum_{j\in\mathcal{T}_m}\gamma_j\left(g_n\left(y,x_j\right)-\bar{g}\left(y\right)\right)\right)^2\right]}{\mu^2} \\
&\leq \lim_{n\to\infty}\sum_{m=\varrho(n)}^{\infty}\bar{\gamma}_m^2 B_g B_T/\mu^2 \\
&= 0.
\end{aligned}$$

C.5 Satisfying Assumption C.1.5

In this section we need to show that for each $\mu > 0$ and for some $T > 0$ we have

$$\lim_{n\to\infty}\Pr\left(\sup_{j\geq n}\max_{0\leq t\leq T}\left\|\sum_{i=m(jT)}^{m(jT+t)-1}\gamma_i\delta M_i\right\|\geq\mu\right)=0. \tag{44}$$

In order to follow the same lines as in Section C.4, we need to show that the second moment of the martingale difference noise, δM_i, is bounded with zero mean. By definition, $\delta M_n\left(\cdot\right)$ has zero mean.

Lemma C.10 *The martingale difference noise, $\delta M_n\left(\cdot\right)$, is bounded in the second moment.*

Proof The claim is immediate from the fact that

$$\mathrm{E}\left[\left(\delta M_n\right)^2\right]=\mathrm{E}\left[\left\|Y_n-g_n\left(y_n,x_n\right)\right\|^2\right]\leq 2\mathrm{E}\left[\left\|Y_n\right\|^2+\left\|g_n\left(y_n,x_n\right)\right\|^2\right],$$

and from Lemma C.3, Lemma C.4, and Lemma C.5. ■

Combining this fact with Lemma C.10, and applying the regenerative decomposition of Section C.4, we conclude that statistically $\delta M_n\left(\cdot\right)$ behaves exactly as $\left(g_n\left(y,x_i\right)-\bar{g}\left(y\right)\right)$ of section C.4 and thus (44) holds.

C.6 Satisfying Assumption C.1.6

In this section we need to prove that there are non-negative measurable functions $\rho_3\left(y\right)$ and $\rho_{n4}\left(x\right)$ such that

$$\left\|g_n\left(y_n,x\right)\right\|\leq\rho_3\left(y_n\right)\rho_{n4}\left(x\right), \tag{45}$$

where $\rho_3\left(y\right)$ is bounded on each bounded y-set, and for each $\mu > 0$ we have

$$\lim_{\tau\to 0}\lim_{n\to\infty}\Pr\left(\sup_{j\geq n}\sum_{i=m(j\tau)}^{m(j\tau+\tau)-1}\gamma_i\rho_{n4}\left(x_i\right)\geq\mu\right)=0.$$

The following lemma states a stronger condition for Assumption C.1.6. In fact, we choose $\rho_3(y)$ to be a positive constant.

Lemma C.11 *If* $\left\|g_n\left(y,x\right)\right\|$ is uniformly bounded for each y, x and n, then *Assumption C.1.6 is satisfied.*

1. If w_n is a martingale sequence then $\Pr\left(\sup_{m\geq 0}\left|w_n\right|\geq\mu\right)\leq\lim_{n\to\infty}\mathrm{E}\left[\left|w_n\right|^2\right]/\mu^2$.

Proof Let us denote the upper bound by the random variable B, i.e.,

$$\|g_n(y,x)\| \leq B, \quad \text{w.p. } 1.$$

Thus

$$\begin{aligned}
\lim_{\tau\to 0}\lim_{n\to\infty}\Pr\left(\sup_{j\geq n}\sum_{i=m(j\tau)}^{m(j\tau+\tau)-1}\gamma_i\rho_{n4}(x_i)\geq\mu\right) &\leq \lim_{\tau\to 0}\lim_{n\to\infty}\Pr\left(\sup_{j\geq n}\sum_{i=m(j\tau)}^{m(j\tau+\tau)-1}\gamma_i B\geq\mu\right)\\
&= \lim_{\tau\to 0}\lim_{n\to\infty}\Pr\left(\sup_{j\geq n}B\sum_{i=m(j\tau)}^{m(j\tau+\tau)-1}\gamma_i\geq\mu\right)\\
&\leq \lim_{\tau\to 0}\Pr(B\tau\geq\mu)\\
&= 0.
\end{aligned}$$

■

Based on Lemma C.11, we are left with proving that $g_n(y,x)$ is uniformly bounded. The following lemma states so.

Lemma C.12 *The function* $g_n(y,x)$ is *uniformly bounded for all* n.

Proof We examine the components of $g_n(y_n,x_n)$. In (35) we showed that

$$g_n(\tilde{\eta}_n,x_n) = \Gamma_\eta(r(x_n)-\tilde{\eta}_n).$$

Since both $r(x_n)$ and $\tilde{\eta}_n$ are bounded by Assumption 3.1 and Lemma C.3 respectively, we have a uniform bound on $g_n(\tilde{\eta}_n,x_n)$. Recalling (36) we have

$$\begin{aligned}
g_n(w_n,x_n) &= \Gamma_w\sum_{k=0}^{\infty}\lambda^k\phi(x_{n-k})\left(r(x_n)-\tilde{\eta}_n+\sum_{y\in\mathcal{X}}P(y|x_n,\theta_n)\phi(y)'w_n-\phi(x_n)'w_n\right)\\
&\leq \Gamma_w\frac{1}{1-\lambda}B_\phi(B_r+B_{\tilde{\eta}}+2B_\phi B_w).
\end{aligned}$$

Finally, recalling (37) we have

$$\begin{aligned}
g_n(\theta_n,x_n) &= \mathrm{E}\left[\tilde{d}(x_n,x_{n+1},w_n)\psi(x_n,u_n,\theta_n)\middle|\mathcal{F}_n\right]\\
&\leq (B_r+B_{\tilde{\eta}}+2B_\phi B_w)B_\psi.
\end{aligned}$$

■

C.7 Satisfying Assumption C.1.7

In this section we show that there are non-negative measurable functions $\rho_1(y)$ and $\rho_{n2}(x)$ such that $\rho_1(y)$ is bounded on each bounded y-set and

$$\|g_n(y_1,x)-g_n(y_2,x)\|\leq\rho_1(y_1-y_2)\rho_{n2}(x) \tag{46}$$

where

$$\lim_{y\to 0}\rho_1(y)=0, \tag{47}$$

and for some $\tau > 0$

$$\Pr\left(\limsup_{j} \sum_{i=j}^{m(t_j+\tau)} \gamma_i \rho_{i2}(x_i) < \infty\right) = 1.$$

From Section C.6 we infer that we can choose $\rho_{n2}(x)$ to be a constant since $g_n(y,x)$ is uniformly bounded. Thus, we need to show the appropriate $\rho_1(\cdot)$ function. The following lemma shows it.

Lemma C.13 *The following functions satisfy (46) and (47).*

1. The function $\rho_1(y) = \|\tilde{\eta}_2 - \tilde{\eta}_1\|$ and $\rho_{n2}(x) = \Gamma_\eta$ for $g_n(\tilde{\eta}, x)$.
   (a) The function $\rho_1(y) = \frac{1}{1-\lambda} B_\phi^2 \left(\sum_{y\in\mathcal{X}} B_w \|P(y|x,\theta_1) - P(y|x,\theta_2)\| + \|w_1 - w_2\|\right)$ and $\rho_{n2}(x) = \Gamma_w$ for $g_n(w,x)$.
   (b) The function $\rho_1(y) = \sum_{y\in\mathcal{X}} B_w \|P(y|x,\theta_1) - P(y|x,\theta_2)\| \cdot B_\psi$ and $\rho_{n2}(x) = 1$ for $g_n(\theta,x)$.

Proof ■

1. Recalling (35) we have for $g_n(\tilde{\eta}, x)$

$$\|g_n(\tilde{\eta}_1, x) - g_n(\tilde{\eta}_2, x)\| \leq \Gamma_\eta \|\tilde{\eta}_2 - \tilde{\eta}_1\|,$$

   thus (46) and (47) are satisfied for 1.

2. Recalling (36) we have for $g_n(w,x)$

$$\begin{aligned}
\|g_n(w_1,x) - g_n(w_2,x)\| &\leq \left\| \Gamma_w \sum_{k}^{\infty} \lambda^k \phi(x_{n-k}) \left(\left(\sum_{y\in\mathcal{X}} P(y|x,\theta_1)\phi(y)' w_1 - \phi(x_n)' w_1 \right) \right.\right. \\
&\quad \left.\left. - \left(\sum_{y\in\mathcal{X}} P(y|x,\theta_2)\phi(y)' w_2 - \phi(x_n)' w_2 \right) \right) \right\| \\
&\leq \frac{\Gamma_w B_\phi^2}{1-\lambda} \left(\sum_{y\in\mathcal{X}} \|P(y|x,\theta_1) w_1 - P(y|x,\theta_2) w_2\| + \|w_1 - w_2\| \right) \\
&\leq \frac{\Gamma_w B_\phi^2}{1-\lambda} \left(\sum_{y\in\mathcal{X}} B_w \|P(y|x,\theta_1) - P(y|x,\theta_2)\| + \|w_1 - w_2\| \right)
\end{aligned}$$

   (a) Trivially, with respect to w (46) and (47) are satisfied. Regarding θ, (46) and (47) are satisfied if we recall the definition of $P(y|x,\theta)$ from (1) and the continuity of $\mu(u|x,\theta)$ from Assumption 3.4.
   (b) Recalling (37) we have for $g_n(\theta,x)$

$$\begin{aligned}
\|g_n(\theta_1,x) - g_n(\theta_2,x)\| &= \left\| \mathrm{E}\left[\tilde{d}(x,y,w_1)\psi(x,u,\theta_1) \middle| \mathcal{F}_n \right] - \mathrm{E}\left[\tilde{d}(x,y,w_2)\psi(x,u,\theta_2) \middle| \mathcal{F}_n \right] \right\| \\
&\leq \sum_{y\in\mathcal{X}} B_w \|P(y|x,\theta_1) - P(y|x,\theta_2)\| \cdot B_\psi.
\end{aligned}$$

   Using similar arguments to 2, (46) and (47) are satisfied for θ.

Appendix D. Proof of Theorem 4.7

In this section we find conditions under which Algorithm 1 converges to a neighborhood of a local maximum. More precisely, we show that $\liminf_{t\to\infty} \|\nabla\eta(\theta(t))\|_2 \leq \epsilon_{\text{app}} + \epsilon_{\text{dyn}}$, where the approximation error, ϵ_{app}, measures the error inherent in the critic's representation, and ϵ_{dyn} is an error related to the single time scale algorithm. We note that the approximation error depends on the basis functions chosen for the critic, and in general can be reduced only by choosing a better representation basis. The term ϵ_{dyn} is the dynamic error, and this error can be reduced by choosing the critic's parameters Γ_η and Γ_w appropriately.

We begin by establishing a variant of Lyapunov's theorem for asymptotic stability[2], where instead of proving asymptotic convergence to a point, we prove convergence to a compact invariant set. Based on this result, we continue by establishing a bound on a time dependent ODE of the first order. This result is used to bound the critic's error in estimating the average reward per stage and the differential values. Finally, using these results, we establish Theorem 4.7.

We denote a closed ball of radius y in some normed vector space, $(\mathbb{R}^L, \|\cdot\|_2)$, by $\mathcal{B}_y$, and its surface by $\partial\mathcal{B}_y$. Also, we denote by $A\backslash B$ a set, which contains all the members of set A which are not members of B. Finally, we define the complement of $\mathcal{B}_y$ by $\mathcal{B}_y^c = \mathbb{R}^L\backslash\mathcal{B}_y$.

The following lemma is similar to Lyapunov's classic theorem for asymptotic stability (Khalil (2002), Theorem 4.1). The main difference is that when the value of the Lyapunov function is unknown inside a ball, convergence can be established to the ball, rather than to a single point.

Lemma D.1 *Consider a dynamical system, $\dot{x} = f(x)$ in a normed vector space, $(\mathbb{R}^L, \|\cdot\|)$, and a closed ball $\mathcal{B}_r \triangleq \{x \,|x \in \mathbb{R}^L, \|x\| \leq r\}$. Suppose that there exists a continuously differentiable scalar function $V(x)$ such that $V(x) > 0$ and $\dot{V}(x) < 0$ for all $x \in \mathcal{B}_r^c$, and $V(x) = 0$ for $x \in \partial\mathcal{B}_r$. Then,*

$$\limsup_{t\to\infty} \|x(t)\| \leq r.$$

Proof We prove two complementary cases. In the first case, we assume that $x(t)$ never enters $\mathcal{B}_r$. On the set $\mathcal{B}_r^c$, $V(x)$ is a strictly decreasing function in t, and it is bounded below, thus it converges. We denote this bound by C, and notice that $C \geq 0$ since for $x \in \mathcal{B}_r^c$, $V(x) > 0$. We prove that $C = 0$ by contradiction. Assume that $C > 0$. Then, $x(t)$ converge to the invariant set $S_C \triangleq \{x|V(x) = C, x \in \mathcal{B}_r^c\}$. For each $x(t) \in \mathcal{S}_C$ we have $\dot{V}(x) < 0$. Thus, $V(x)$ continues to decrease which contradicts the boundedness from below. As a result, $V(x(t)) \to 0$.

In the second case, let us suppose that at some time, denoted by t_0, $x(t_0) \in \mathcal{B}_r$. We argue that the trajectory never leaves $\mathcal{B}_r$. Let us assume that at some time t_2, the trajectory $x(t)$ enters the set $\partial\mathcal{B}_{r+\epsilon}$. Then on this set, we have $V(x(t_2)) > 0$. By the continuity of the trajectory $x(t)$, the trajectory must go through the set $\partial\mathcal{B}_r$. Denote the hitting time of this set by t_1. By definition we have $V(x(t_1)) = 0$. Without loss of generality, we assume that the trajectory in the times $t_1 < t \leq t_2$ is restricted to the set $\mathcal{B}_{r+\epsilon}/\mathcal{B}_r$. Thus, since $\dot{V}(x(t)) \leq 0$ for $x \in \mathcal{B}_{r+\epsilon}/\mathcal{B}_r$ we have

$$V(x(t_2)) = V(x(t_1)) + \int_{t_1}^{t_2} \dot{V}(x(t))dt < V(x(t_1)),$$

which contradicts the fact that $V(x(t_2)) \geq V(x(t_1))$. Since this argument holds for all $\epsilon > 0$, the trajectory $x(t)$ never leaves $\mathcal{B}_r$. ■

The following lemma will be applied later to the linear equations (27), and more specifically, to the ODEs describing the dynamics of $\tilde{\eta}$ and w. It bounds the difference between an ODE's state variables and some time dependent functions.

2. We say that the equilibrium point $x = 0$ of the system $\dot{x}$=f(x) is *stable* if for each $\epsilon > 0$ there exists a $\delta > 0$ such that $\|x(0)\| < \delta \Rightarrow \|x(t)\| < \epsilon$ for all $t \geq 0$. We say that the point $x = 0$ is *asymptotically stable* if it is stable and there exists a $\delta > 0$ such that $\|x(0)\| < \delta$ implies $\lim_{t\to\infty} x(t) = 0$ (see Khalil (2002) for more details).

Lemma D.2 *Consider the following ODE in a normed space* $(\mathbb{R}^L, \|\cdot\|_2)$

$$\begin{cases} \frac{d}{dt}X(t) = \mathcal{M}(t)\left(X(t) - F_1(t)\right) + F_2(t), \\ X(0) = X_0, \end{cases} \tag{48}$$

where for sufficiently large t *.*

1. $\mathcal{M}(t) \in \mathbb{R}^{L\times L}$ *is a continuous matrix which satisfies* $\max_{\|x\|=1} x'\mathcal{M}(t)\,x \le -\gamma < 0$ *for* $t \in \mathbb{R}$,
2. $F_1(t) \in \mathbb{R}^L$ *satisfies* $\|dF_1(t)/dt\|_2 \le B_{F1}$,
3. $F_2(t) \in \mathbb{R}^L$ *satisfies* $\|F_2(t)\|_2 \le B_{F2}$.

Then, the solution of the ODE satisfies $\limsup_{t\to 0} \|X(t) - F_1(t)\|_2 \le (B_{F1} + B_{F2})/\gamma$.

Proof We express (48) as

$$\frac{d}{dt}\left(X(t) - F_1(t)\right) = \mathcal{M}(t)\left(X(t) - F_1(t)\right) - \frac{d}{dt}F_1(t) + F_2(t), \tag{49}$$

and define

$$Z(t) \triangleq \left(X(t) - F_1(t)\right), \quad G(t) \triangleq -\frac{d}{dt}F_1(t) + F_2(t).$$

Therefore, (49) can be written as

$$\dot{Z}(t) = \mathcal{M}(t)Z(t) + G(t),$$

where $\|G(t)\| \le B_G \triangleq B_{F1} + B_{F2}$. In view of Lemma D.1, we consider the function

$$V(Z) = \frac{1}{2}\left(\|Z(t)\|_2^2 - B_G^2/\gamma^2\right).$$

Let $\mathcal{B}_r$ be a ball with a radius $r = B_G/\gamma$. Thus we have $V(Z) > 0$ for $Z \in \mathcal{B}_r^c$ and $V(Z) = 0$ for $X \in \partial\mathcal{B}_r$. In order to satisfy the assumptions of Lemma D.1 the condition that $\dot{V}(Z) < 0$ needs to be verified. For $\|Z(t)\|_2 > B_G/\gamma$ we have

$$\begin{aligned} \dot{V}(Z) &= \left(\nabla_X V\right)' \dot{Z}(t) \\ &= Z(t)'\mathcal{M}(t)Z(t) + Z(t)'G(t) \\ &= \|Z(t)\|_2^2 \frac{Z(t)'}{\|Z(t)\|_2}\mathcal{M}(t)\frac{Z(t)}{\|Z(t)\|_2} + Z(t)'G(t) \\ &\le \|Z(t)\|_2^2 \max_{\|Y(t)\|_2=1} Y(t)'\mathcal{M}(t)Y(t) + \|Z(t)\|_2 \|G(t)\|_2 \\ &= \|Z(t)\|_2 \left(-\gamma\|Z(t)\|_2 + B_G\right) \\ &< 0. \end{aligned}$$

As a result, the assumptions of Lemma D.1 are valid and the Lemma is proved. ■

The following lemma shows that the matrix $A(\theta)$, defined in (26), satisfies the conditions of Lemma D.2. For the following lemmas, we define the weighted norm $\|w\|_{\Pi(\theta)}^2 \triangleq \|w'\Pi(\theta)\,w\|_2$.

Lemma D.3 *The following inequalities hold:*

1. For any $w \in \mathbb{R}^L$and for all $\theta \in \mathbb{R}^K$, $\|P(\theta)\,w\|_{\Pi(\theta)} < \|w\|_{\Pi(\theta)}$..

(a) The matrix $M(\theta)$ satisfies $\|M(\theta)w\|_{\Pi(\theta)} < \|w\|_{\Pi(\theta)}$ for all $\theta \in \mathbb{R}^K$ and $w \in \mathbb{R}^L$.

(b) The matrix $\Pi(\theta)(M(\theta) - I)$ satisfies $x'\Pi(\theta)(M(\theta) - I)x < 0$ for all $x \in \mathbb{R}^L$ and for all $\theta \in \mathbb{R}^K$.

(c) There exists a positive scalar γ such that $w'A(\theta)w < -\gamma$ for all $w'w = 1$.

Proof The following proof is similar in many aspects to the proof of Lemma 6.6 of Bertsekas and Tsitsiklis (1996). ■

1. By using Jensen's inequality for the function $f(\alpha) = \alpha^2$ we have

$$\left(\sum_{y\in\mathcal{X}} P(y|x,\theta)\, w(y)\right)^2 \leq \sum_{y\in\mathcal{X}} P(y|x,\theta)\, w(y)^2, \quad \forall x \in \mathcal{X}. \tag{50}$$

If in Jensen's inequality we have a strictly convex fiction and non-degenerate probability measures then the inequality is strict. The function $f(\alpha)$ is strictly convex, and by Assumption 3.2 the matrix $P(\theta)$ is aperiodic, which implies that the matrix $P(\theta)$ is not a permutation matrix. As a result, there exists $x_0 \in \mathcal{X}$ such that the probability measure $P(y|x_0,\theta)$ is not degenerate, thus, the inequality in (50) is strict, i.e.,

$$\left(\sum_{y\in\mathcal{X}} P(y|x_0,\theta)\, w(y)\right)^2 < \sum_{y\in\mathcal{X}} P(y|x_0,\theta)\, w(y)^2. \tag{51}$$

Then, we have

$$\begin{aligned}
\|P(\theta)w\|_{\Pi(\theta)} &= w'P(\theta)'\Pi(\theta)P(\theta)w \\
&= \sum_{x\in\mathcal{X}} \pi(x|\theta)\left(\sum_{y\in\mathcal{X}} P(y|x,\theta)\, w(y)\right)^2 \\
&< \sum_{x\in\mathcal{X}} \pi(x|\theta)\sum_{y\in\mathcal{X}} P(y|x,\theta)\, w(y)^2 \\
&= \sum_{y\in\mathcal{X}} w(y)^2 \sum_{x\in\mathcal{X}} \pi(x|\theta)\, P(y|x,\theta) \\
&= \sum_{y\in\mathcal{X}} w(y)^2\, \pi(y|\theta) \\
&= \|w\|_{\Pi(\theta)},
\end{aligned}$$

where in the inequality we have used (51).

(a) Using the triangle inequality and 1 we have

$$\begin{aligned}
\|M(\theta)w\|_{\Pi(\theta)} &= \left\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m P(\theta)^{m+1} w\right\|_{\Pi(\theta)} \\
&\leq (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \left\|P(\theta)^{m+1} w\right\|_{\Pi(\theta)} \\
&< (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \|w\|_{\Pi(\theta)} \\
&= \|w\|_{\Pi(\theta)}.
\end{aligned}$$

(b) By definition

$$\begin{aligned} x'\Pi(\theta)M(\theta)x &= x'\Pi(\theta)^{1/2}\Pi(\theta)^{1/2}M(\theta)x \\ &\leq \left\|\Pi(\theta)^{1/2}x\right\| \cdot \left\|\Pi(\theta)^{1/2}M(\theta)x\right\| \\ &= \|x\|_{\Pi(\theta)} \|M(\theta)x\|_{\Pi(\theta)} \\ &< \|x\|_{\Pi(\theta)} \|x\|_{\Pi(\theta)} . \\ &= x'\Pi(\theta)x, \end{aligned}$$

where in the first inequality we have used the Cauchy-Schwartz inequality, and in the second inequality we have used 1. Thus, $x'\Pi(\theta)(M(\theta)-I)x < 0$ for all $x \in \mathbb{R}$, which implies that $\Pi(\theta)(M(\theta)-I)$ is a negative definite (ND) matrix[3].

(c) From 3, we know that for all $\theta \in \mathbb{R}^K$ and all $w \in \mathbb{R}^{|\mathcal{X}|}$ satisfying $w'w = 1$, we have $w'\Pi(\theta)(M(\theta)-I)w < 0$, and by Assumption (3.2), this is true also for the closure of $\left\{\Pi(\theta)(M(\theta)-I)\,|\theta \in \mathbb{R}^K\right\}$. Thus, there exists a positive scalar, γ', satisfying

$$w'\Pi(\theta)(M(\theta)-I)w \leq -\gamma' < 0.$$

By Assumption 4.3 the rank of the matrix Φ is full, thus there exists a scalar γ such that for all $w \in \mathbb{R}^L$, where $w'w = 1$, we have $w'A(\theta)w \leq -\gamma < 0$.

The following Lemma establishes the boundedness of $\dot{\theta}$.

Lemma D.4 *There exists a constant* $B_{\theta 1} \triangleq B_{\eta 1} + B_\psi(B_D + B_r + B_{\tilde{\eta}} + 2B_\phi B_w)$ *such that* $\|\dot{\theta}\|_2 \leq B_{\theta 1}$.

Proof Recalling (27)

$$\begin{aligned} \left\|\dot{\theta}\right\|_2 &= \left\|\nabla_\theta \eta(\theta) + \sum_{x,y\in\mathcal{X}\times\mathcal{X},u\in\mathcal{U}} D^{(x,u,y)}(\theta)\left(d(x,y,\theta) - \tilde{d}(x,y,w)\right)\right\|_2 \\ &\leq B_{\eta 1} + \sum_{x,y\in\mathcal{X}\times\mathcal{X},u\in\mathcal{U}} \left\|D^{(x,u,y)}(\theta)\right\|_2 \left\|d(x,y,\theta) - \tilde{d}(x,y,w)\right\|_2 \\ &\leq B_{\eta 1} + B_\psi(B_D + B_r + B_{\tilde{\eta}} + 2B_\phi B_w) \\ &\triangleq B_{\theta 1}. \end{aligned}$$

■

Based on Lemma (D.4), the following Lemma shows the boundedness of $(\eta(\theta(t)) - \tilde{\eta})$.

Lemma D.5 *We have*

$$\limsup_{t\to\infty} |\eta(\theta(t)) - \tilde{\eta}| \leq \frac{B_{\Delta\eta}}{\Gamma_\eta},$$

where $B_{\Delta\eta} \triangleq B_{\eta_1} B_{\theta 1}$.

Proof Using the Cauchy-Schwartz inequality we have

$$\begin{aligned} |\dot{\eta}(\theta)| &= |\nabla\eta(\theta)'\dot{\theta}| \\ &\leq \|\nabla\eta(\theta)\|_2 \|\dot{\theta}\|_2 \\ &\leq B_{\eta_1} B_{\theta 1}. \end{aligned} \tag{52}$$

3. Usually, a ND matrix is defined for Hermitian matrices, i.e., if B is an Hermitian matrix and it satisfies $x'Bx < 0$ for all $x \in \mathbb{C}^K$ then B is a NSD matrix . We use here a different definition which states that a square matrix B is a ND matrix if it is real and it satisfies $x'Bx < 0$ for all $x \in \mathbb{R}^k$ (see Horn and Johnson (1985) p. 399).

Recalling the equation for $\tilde{\eta}$ in (27) we have

$$\dot{\tilde{\eta}} = \Gamma_\eta \left(\eta(\theta) - \tilde{\eta}\right).$$

We conclude by applying Lemma D.2 and using (52) that

$$\limsup_{t\to\infty} |\eta(\theta(t)) - \tilde{\eta}| \leq \frac{B_{\eta_1} B_{\theta 1}}{\Gamma_\eta} = \frac{B_{\Delta\eta}}{\Gamma_\eta}. \tag{53}$$

■

In (53) we see that the bound on $|\eta(\theta) - \tilde{\eta}|$ is controlled by Γ_η, where larger values of Γ_η ensure smaller values of $|\eta(\theta) - \tilde{\eta}|$. Next, we bound $\|w^*(\theta) - w\|_2$. We recall the second equation of (27)

$$\begin{aligned}
\dot{w} &= \Psi_w \left[\Gamma_w \left(A\left(\theta\right) w + b\left(\theta\right) + G(\theta)(\eta(\theta) - \tilde{\eta})\right)\right], \\
A\left(\theta\right) &= \Phi' \Pi\left(\theta\right)\left(M - I\right)\Phi, \\
M\left(\theta\right) &= \left(1-\lambda\right)\sum_{m=0}^{\infty} \lambda^m P\left(\theta\right)^{m+1}, \\
b\left(\theta\right) &= \Phi' \Pi\left(\theta\right)\sum_{m=0}^{\infty} \lambda^m P\left(\theta\right)^{m}\left(r - \eta\left(\theta\right)\right), \\
G(\theta) &= \Phi' \Pi\left(\theta\right)\sum_{m=0}^{\infty} \lambda^m P\left(\theta\right)^{m}.
\end{aligned}$$

We can write the equation for $\dot{w}$ as

$$\dot{w} = \Psi_w \left[\Gamma_w \left(A\left(\theta\right)\left(w - w^*\left(\theta\right)\right) + G(\theta)(\eta(\theta) - \tilde{\eta})\right)\right],$$

where $w^* = -A\left(\theta\right)^{-1} b\left(\theta\right)$. In order to use Lemma D.2, we need to demonstrate the boundedness of $\left\|\frac{d}{dt} w^*\right\|$. The following lemma does so.

Lemma D.6

1. There exists a positive constant, $B_b \triangleq \frac{1}{1-\lambda}\left|\mathcal{X}\right|^3 L B_\Phi B_r$, such that $\left\|b\left(\theta\right)\right\|_2 \leq B_b$.
   (a) There exists a positive constant, $B_G \triangleq \frac{1}{1-\lambda}\left|\mathcal{X}\right|^3 L B_\Phi$, such that $\left\|G\left(\theta\right)\right\|_2 \leq B_G$.
   (b) There exist positive constants, $\tilde{B} = B_{\pi 1}\left(B_r + B_\eta\right) B_{\theta 1} + B_{P1}\left(B_r + B_\eta\right) B_{\theta 1} + B_{\eta 1} B_{\theta 1}$ and $B_{b1} \triangleq \frac{1}{1-\lambda}\left|\mathcal{X}\right|^3 B_\Phi B_r \tilde{B}$, such that we have $\left\|\dot{b}\left(\theta\right)\right\|_2 \leq B_{b1}$.
   (c) There exist constants b_A and B_A such that
   $$0 < b_A \leq \left\|A\left(\theta\right)\right\|_2 \leq B_A.$$
   (d) There exist a constants B_{A1} such that
   $$\left\|A\left(\theta\right)\right\|_2 \leq B_{A1}.$$
   (e) We have
   $$\left\|\frac{d}{dt}\left(A\left(\theta\right)^{-1}\right)\right\|_2 \leq b_A^2 B_{A1}.$$

(f) There exists a positive constant, B_{w1}, such that

$$\left\| \frac{d}{dt} w^* \right\|_2 \leq B_{w1}.$$

Proof ■

1. We show that the entries of the vector $b(\theta)$ are uniformly bounded in θ, therefore, its norm is uniformly bounded in θ. Let us look at the i-th entry of the vector $b(\theta)$ (we denote by $[\cdot]_j$ the j-th row of a matrix or a vector)

$$\begin{aligned} |[b(\theta)]_i| &= \left| \left[\Phi' \Pi(\theta) \sum_{m=0}^{\infty} \lambda^m P(\theta)^m (r - \eta(\theta)) \right]_i \right| \\ &\leq \sum_{m=0}^{\infty} \lambda^m \left| [\Phi' \Pi(\theta) P(\theta)^m (r - \eta(\theta))]_i \right| \\ &= \sum_{m=0}^{\infty} \lambda^m \left| \sum_{l=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{X}|} \sum_{k=1}^{|\mathcal{X}|} [\Phi']_{ik} \Pi_{kj}(\theta) [P(\theta)^m]_{jl} (r_l - \eta(\theta)) \right| \\ &\leq \frac{1}{1-\lambda} |\mathcal{X}|^3 B_\Phi B_r, \end{aligned}$$

thus $\|b(\theta)\|_2 \leq \frac{1}{1-\lambda} |\mathcal{X}|^3 L B_\Phi B_r$ is uniformly bounded in θ.

2. The proof is accomplished by similar argument to section 1.

3. Similarly to section 1, we show that the entries of the vector $\dot{b}(\theta)$ are uniformly bounded in θ, therefore, its norm is uniformly bounded in θ. First, we show that the following function of $\theta(t)$ is bounded.

$$\begin{aligned} \left| \frac{d}{dt} \left(\Pi_{kj}(\theta) [P(\theta)^m]_{jl} (r_l - \eta(\theta)) \right) \right| &= \left| \nabla_\theta \left(\Pi_{kj}(\theta) [P(\theta)^m]_{jl} (r_l - \eta(\theta)) \right) \dot{\theta} \right| \\ &\leq \left| (\nabla_\theta \Pi_{kj}(\theta)) [P(\theta)^m]_{jl} (r_l - \eta(\theta)) \dot{\theta} \right| \\ &\quad + \left| \Pi_{kj}(\theta) [\nabla_\theta P(\theta)^m]_{jl} (r_l - \eta(\theta)) \dot{\theta} \right| \\ &\quad + \left| \Pi_{kj}(\theta) [P(\theta)^m]_{jl} \nabla_\theta (r_l - \eta(\theta)) \dot{\theta} \right| \\ &\leq B_{\pi 1} (B_r + B_\eta) \cdot B_{\theta 1} + B_{P1} (B_r + B_\eta) B_{\theta 1} + B_{\eta 1} B_{\theta 1} \\ &= \tilde{B}, \end{aligned}$$

where we used the triangle and Cauchy-Schwartz inequalities in the first and second inequalities respectively, and Lemmas 3.6 and D.4 in the second inequality. Thus,

$$\begin{aligned}
\left|\left[\dot{b}(\theta)\right]_i\right| &= \left|\left[\Phi'\Pi(\theta)\sum_{m=0}^{\infty}\lambda^m P(\theta)^m (r-\eta(\theta))\right]_i\right| \\
&\leq \sum_{m=0}^{\infty}\lambda^m \left|\left[\Phi'\Pi(\theta) P(\theta)^m (r-\eta(\theta))\right]_i\right| \\
&= \sum_{m=0}^{\infty}\lambda^m \left|\sum_{l=1}^{|\mathcal{X}|}\sum_{j=1}^{|\mathcal{X}|}\sum_{k=1}^{|\mathcal{X}|}[\Phi']_{ik}\frac{d}{dt}\left(\Pi_{kj}(\theta)\left[P(\theta)^m\right]_{jl}(r_l-\eta(\theta))\right)\right| \\
&\leq \frac{1}{1-\lambda}|\mathcal{X}|^3 B_\Phi B_r \tilde{B} \\
&= B_{b1}.
\end{aligned}$$

4. Since $A(\theta)$ satisfies $y'A(\theta)y < 0$ for all nonzero y, it follows that all its eigenvalues are nonzero. Therefore, the eigenvalues of $A(\theta)' A(\theta)$ are all positive and real since $A(\theta)' A(\theta)$ is a symmetric matrix. Since by Assumption 3.2 this holds for all $\theta \in \mathbb{R}^K$, there is a global minimum, b_A, and a global maximum, B_A, such that

$$B_A^2 \geq \lambda_{\max}\left(A(\theta)' A(\theta)\right) \geq \lambda_{\min}\left(A(\theta)' A(\theta)\right) \geq b_A^2, \quad \forall\theta \in \mathbb{R}^K,$$

where we denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the minimal and maximal eigenvalues of the matrix respectively. Using Horn and Johnson (1985) section 5.6.6, we have $\lambda_{max}\left(A(\theta)' A(\theta)\right) = \|A(\theta)\|_2$, thus, we get an upper bound on the matrix norm. Let us look at the norm of $\left\|A(\theta)^{-1}\right\|_2$,

$$\begin{aligned}
\left\|A(\theta)^{-1}\right\|_2^2 &= \lambda_{\max}\left(\left(A(\theta)^{-1}\right)' A(\theta)^{-1}\right) \\
&= \lambda_{\max}\left(\left(A(\theta)'\right)^{-1} A(\theta)^{-1}\right) \\
&= \lambda_{\max}\left(\left(A(\theta) A(\theta)'\right)^{-1}\right) \\
&= 1/\lambda_{\min}\left(A(\theta) A(\theta)'\right) \\
&= 1/\lambda_{\min}\left(\left(A(\theta)' A(\theta)\right)'\right) \\
&= 1/\lambda_{\min}\left(A(\theta)' A(\theta)\right),
\end{aligned}$$

thus, we the lower bound on $\left\|A(\theta)^{-1}\right\|_2$ is $\sqrt{1/\lambda_{\min}\left(A(\theta)' A(\theta)\right)}$, i.e., b_A.

5. Let us look at the ij entry of the matrix $\frac{d}{dt}A(\theta)$, where using similar arguments to section 2 we get

$$\begin{aligned}\left[\left|\frac{d}{dt}A(\theta)\right|\right]_{ij} &= \left[\left|\frac{d}{dt}\left(\Phi'\Pi(\theta)\left((1-\lambda)\sum_{m=0}^{\infty}\lambda^m P(\theta)^{m+1}-I\right)\Phi\right)\right|\right]_{ij}\\ &\leq \left[\left|\Phi'\frac{d}{dt}(\Pi(\theta))\left((1-\lambda)\sum_{m=0}^{\infty}\lambda^m P(\theta)^{m+1}-I\right)\Phi\right|\right]_{ij}\\ &\quad+\left[\left|\Phi'\Pi(\theta)\frac{d}{dt}\left((1-\lambda)\sum_{m=0}^{\infty}\lambda^m P(\theta)^{m+1}-I\right)\Phi\right|\right]_{ij}\\ &\leq B_\Phi B_{\pi 1}\frac{1}{1-\lambda}B_\Phi+B_\Phi\frac{1}{(1-\lambda)^2}B_{P1}B_\Phi.\end{aligned}$$

Since the matrix entries are uniformly bounded in θ, so is the matrix $\frac{d}{dt}A(\theta)'\frac{d}{dt}A(\theta)$, and so is the largest eigenvalue of $\frac{d}{dt}A(\theta)'\frac{d}{dt}A(\theta)$ which implies the uniform boundedness of $\left\|\frac{d}{dt}A(\theta)\right\|_2$.

6. For a general invertible square matrix, $X(t)$, we have

$$0 = \frac{d}{dt}I=\frac{d}{dt}\left(X(t)^{-1}X(t)\right)=\frac{d}{dt}\left(X(t)^{-1}\right)X(t)+X(t)^{-1}\frac{d}{dt}(X(t)).$$

Rearranging it we get

$$\frac{d}{dt}\left(X(t)^{-1}\right)=-X(t)^{-1}\frac{d}{dt}(X(t))X(t)^{-1}.$$

Using this identity yields

$$\begin{aligned}\left\|\frac{d}{dt}\left(A(\theta)^{-1}\right)\right\|_2 &= \left\|-A(\theta)^{-1}\frac{d}{dt}(A(\theta))A(\theta)^{-1}\right\|_2\\ &\leq \left\|A(\theta)^{-1}\right\|_2\cdot\left\|\frac{d}{dt}(A(\theta))\right\|_2\cdot\left\|-A(\theta)^{-1}\right\|_2\\ &= b_A^2B_{A1}.\end{aligned}$$

7. Examining the norm of $\frac{d}{dt}w^*$ yields

$$\begin{aligned}\left\|\frac{d}{dt}w^*\right\|_2 &= \left\|\frac{d}{dt}\left(A(\theta)^{-1}b(\theta)\right)\right\|_2\\ &= \left\|\frac{d}{dt}A(\theta)^{-1}b(\theta)+A(\theta)^{-1}\frac{d}{dt}b(\theta)\right\|_2\\ &\leq b_A^2B_{A1}\frac{1}{1-\lambda}|\mathcal{X}|^3B_\Phi B_r+b_A\tilde{B}\\ &= B_{w1}.\end{aligned}$$

We wish to use Lemma D.2 for (27), thus, we show that the assumptions of Lemma D.2 are valid.

Lemma D.7

1. We have

$$\limsup_{t\to\infty}\|w^*(\theta(t))-w(t)\|_2\leq\frac{1}{\Gamma_w}B_{\Delta w},\tag{54}$$

where

$$B_{\Delta w}\triangleq\frac{B_{w1}+B_G\frac{B_{\Delta\eta}}{\Gamma_\eta}}{\gamma}.$$

(a) We have

$$\limsup_{t\to\infty} \|h(\theta(t)) - \tilde{h}(w(t))\|_\infty \leq \frac{B_{\Delta h1}}{\Gamma_w} + \frac{\epsilon_{\text{app}}}{\sqrt{b_\pi}},$$

where

$$B_{\Delta h} \triangleq |\mathcal{X}| L \left(B_{\Delta w}\right)^2.$$

Proof ■

1. Without loss of generality, we can eliminate the projection operator since we can choose B_w to be large enough such that $w^*(\theta)$ will be inside the bounded space. We take $\mathcal{M}(t) = A(\theta)$, $F_1(t) = w^*(\theta(t))$, and $F_2(t) = G(\theta)(\eta(\theta) - \tilde{\eta})$. By previous lemmas we can see that the Assumption D.2 holds. By Lemma D.6 (6), $\|\dot{w}^*(\theta)\|_2$ is bounded by B_{w_1}, by Lemma D.5 we have a bound on $|(\eta(\theta) - \tilde{\eta})|$, and by Lemma D.3 we have a bound on $w' A(\theta) w$. Using these bounds and applying Lemma D.2 provides the desired result.

   (a) Suppressing the time dependence for simplicity and expressing $\|h(\theta) - \tilde{h}(w)\|_\infty$ using ϵ_{app} and the previous result yields

$$\begin{aligned} \|h(\theta) - \tilde{h}(w)\|_\infty &\leq \|h(\theta) - \tilde{h}(w)\|_2 \\ &= \|h(\theta) - \tilde{h}(w^*) + \tilde{h}(w^*) - \tilde{h}(w)\|_2 \\ &\leq \|h(\theta) - \tilde{h}(w^*)\|_2 + \|\tilde{h}(w^*) - \tilde{h}(w)\|_2 \end{aligned} \tag{55}$$

   For the first term on the r.h.s. of the final equation in (55) we have

$$\begin{aligned} \|h(\theta) - \tilde{h}(w^*)\|_2 &= \left\| \left(\Pi(\theta)^{-\frac{1}{2}}\right) \left(\Pi(\theta)^{\frac{1}{2}}\right) \left(h(\theta) - \tilde{h}(w^*)\right) \right\|_2 \\ &\leq \left\|\Pi(\theta)^{-\frac{1}{2}}\right\|_2 \left\|h(\theta) - \tilde{h}(w^*)\right\|_{\Pi(\theta)} \\ &\leq \frac{\epsilon_{\text{app}}}{(b_\pi)^{\frac{1}{2}}} \end{aligned} \tag{56}$$

   where we use the sub-additivity of the matrix norms in the first inequality, and Lemma 3.6 and the (17) in the last inequality. For the second term on the r.h.s. of the final equation in (55) we have

$$\begin{aligned} \|\tilde{h}(w^*) - \tilde{h}(w)\|_2^2 &= \|\Phi(w^*(\theta) - w)\|_2^2 \\ &= \sum_{k=1}^{|\mathcal{X}|} \left(\sum_{l=1}^{L} \phi_l(k) \left(w_l^*(\theta) - w_l\right) \right)^2 \\ &\leq \sum_{k=1}^{|\mathcal{X}|} \left(\left(\sum_{l=1}^{L} \phi_l^2(k) \right)^{\frac{1}{2}} \left(\sum_{l=1}^{L} \left(w_l^*(\theta) - w_l\right)^2 \right)^{\frac{1}{2}} \right)^2 \\ &\leq \sum_{k=1}^{|\mathcal{X}|} \left(\sum_{l=1}^{L} \phi_l^2(k) \right) \left(\sum_{l=1}^{L} \left(w_l^*(\theta) - w_l\right)^2 \right) \\ &\leq |\mathcal{X}| L \|w^*(\theta) - w\|_2^2 \\ &= |\mathcal{X}| L \left(B_{\Delta w}\right)^2. \end{aligned} \tag{57}$$

   Combining (54)-(57) yields the desired result.

Using Lemma D.7 we can provide a bound on second term of (27).

Lemma D.8 *We have*

$$\limsup_{t\to\infty}\left\|\sum_{x,y\in\mathcal{X}\times\mathcal{X},u\in\mathcal{U}} D^{(x,u,y)}(\theta)\left(d(x,y,\theta)-\tilde{d}(x,y,w)\right)\right\|_2 \leq \frac{B_{\Delta td1}}{\Gamma_w}+\frac{B_{\Delta td2}}{\Gamma_\eta}+B_{\Delta td3}\epsilon_{app}$$

where

$$B_{\Delta td1}=\frac{1}{\Gamma_w}\cdot 2B_\Psi B_{\Delta h1},\quad B_{\Delta td2}=\frac{1}{\Gamma_\eta}\cdot B_{\Delta\eta}B_\Psi,\quad B_{\Delta td3}=\frac{2B_\Psi}{\sqrt{b_\pi}}.$$

Proof Simplifying the notation by suppressing the time dependence, we bound the TD signal in the limit, i.e.,

$$\begin{aligned}\limsup_{t\to\infty}|d(x,y,\theta)-\tilde{d}(x,y,w)| &= \limsup_{t\to\infty}\left|(r(x)-\eta(\theta)+h(y,\theta)-h(x,\theta))-\left(r(x)-\tilde{\eta}+\tilde{h}(y,w)-\tilde{h}(x,w)\right)\right|\\ &\leq \limsup_{t\to\infty}|\eta(\theta)-\tilde{\eta}|+\limsup_{t\to\infty}2\left\|h(\theta)-\tilde{h}(w)\right\|_\infty\\ &= \frac{B_{\Delta\eta}}{\Gamma_\eta}+2\left(\frac{B_{\Delta h1}}{\Gamma_w}+\frac{\epsilon_{\text{app}}}{\sqrt{b_\pi}}\right).\end{aligned}$$

With some more algebra we have

$$\begin{aligned}&\limsup_{t\to\infty}\left\|\sum_{x,y\in\mathcal{X}\times\mathcal{X},u\in\mathcal{U}} D^{(x,u,y)}(\theta)\left(d(x,y,\theta)-\tilde{d}(x,y,w)\right)\right\|\\ &\leq \limsup_{t\to\infty}\sum_{x,y\in\mathcal{X}\times\mathcal{X},u\in\mathcal{U}}\pi(x)\,P(u|x,\theta_n)\,P(y|x,u)\,\|\psi(x,u,\theta_n)\|\cdot\left|d(x,y,\theta)-\tilde{d}(x,y,w)\right|\\ &\leq B_\Psi\left(\frac{B_{\Delta\eta}}{\Gamma_\eta}+2\left(\frac{B_{\Delta h1}}{\Gamma_w}+\frac{\epsilon_{\text{app}}}{\sqrt{b_\pi}}\right)\right)\\ &= \frac{B_{\Delta td1}}{\Gamma_w}+\frac{B_{\Delta td2}}{\Gamma_\eta}+B_{\Delta td3}\epsilon_{\text{app}}.\end{aligned}$$

■

We see that the term in this bound is adjustable by choosing appropriate Γ_η and Γ_w. The concluding lemma proves the conclusion of Theorem 4.7.

Proof of Theorem 4.7

We define

$$B_{\nabla\eta}\triangleq\frac{B_{\Delta td1}}{\Gamma_w}+\frac{B_{\Delta td2}}{\Gamma_\eta}+B_{\Delta td3}\epsilon_{\text{app}}.$$

For an arbitrary $\delta>0$, define the set

$$\mathcal{B}_\delta\triangleq\{\theta:\ \|\nabla\eta(\theta)\|\leq B_{\nabla\eta}+\delta\}. \tag{58}$$

We claim that the trajectory $\eta(\theta)$ visits $\mathcal{B}_\delta$ infinitely often. Assume the contrary that

$$\liminf_{t\to\infty}\|\nabla\eta(\theta)\|_2 > B_{\nabla\eta}+\delta. \tag{59}$$

Thus, on the set $\mathcal{B}_\delta^c$ for t large enough we have

$$\begin{aligned}
\dot{\eta}(\theta) &= \nabla\eta(\theta)\cdot\dot{\theta} \\
&= \nabla\eta(\theta)\cdot\left(\nabla\eta(\theta)+\sum_{x,y\in\mathcal{X}\times\mathcal{X}} D^{(x,y)}(\theta)\left(d(x,y)-\tilde{d}(x,y)\right)\right) \\
&= \|\nabla\eta(\theta)\|_2^2+\nabla\eta(\theta)\cdot\left(\sum_{x,y\in\mathcal{X}\times\mathcal{X}} D^{(x,y)}(\theta)\left(d(x,y)-\tilde{d}(x,y)\right)\right) \\
&\geq \|\nabla\eta(\theta)\|_2^2-\|\nabla\eta(\theta)\|_2\left\|\sum_{x,y\in\mathcal{X}\times\mathcal{X}} D^{(x,y)}(\theta)\left(d(x,y)-\tilde{d}(x,y)\right)\right\|_2 \\
&= \|\nabla\eta(\theta)\|_2\left(\|\nabla\eta(\theta)\|_2-B_{\nabla\eta}\right) \\
&\geq \|\nabla\eta(\theta)\|_2\left(B_{\nabla\eta}+\delta-B_{\nabla\eta}\right) \\
&> (B_{\nabla\eta}+\delta)\delta.
\end{aligned} \tag{60}$$

By (59), there exists a time t_0 which for all $t > t_0$ we have $\eta(\theta) \in \mathcal{B}_\delta^c$. Therefore,

$$\eta(\infty) = \eta(t_0) + \int_{t_0}^{\infty} \dot{\eta}(\theta)dt > \eta(t_0) + \int_{t_0}^{\infty} (B_D+\delta)\delta dt = \infty, \tag{61}$$

which contradicts the boundedness of $\eta(\theta)$. Since the claim holds for all $\delta > 0$, the result follows.

References

T. Archibald, K. McKinnon, and L. Thomas. On the generation of markov decision processes. *Journal of the Operational Research Society*, 1995.

D. Baras and R. Meir. Reinforcement learning, spike time dependent plasticity and the bcm rule. *Neural Comput.*, 19(8):2245–2279, Aug 2007.

J. Baxter and P. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

J. Baxter, P. Bartlett, and E. Greensmith. Variance reduction techniques for gradient estimates in reinforcement learning. 2004.

D. Bertsekas. *Dynamic Programming and Optimal Control, Vol I & II, 3rd Ed.* Athena Scinetific, 2006.

D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic Programming.* Athena Scinetific, 1996.

S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Incremental natural actor-critic algorithms. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 105–112, Cambridge, MA, 2008a. MIT Press.

S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor-critic algorithms. *Automatica*, 2008b. In press.

V. Borkar. Stochastic approximation with two time scales. *Syst. Control Lett.*, 29(5):291–294, 1997.

P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer, 1999.

X. Cao. Stochastic learning and optimization: A sensitivity-based approach (international series on discrete event dynamic systems). 2007.

X. Cao and H. Chen. Pertubation realization, potentials, and sensitivity analysis of markov processes. *IEEE Trans. Automat. Contr*, 42:1382–1393, 1997.

N. Daw, Y. Niv, , and P. Dayan. *Actions, Policies, Values, and the Basal Ganglia - In: Bezard, E editor, Recent Breakthroughs in Basal Ganglia Research*. Nova Science Publishers Inc., 2006.

D. DiCastro, D. Volkinstein, and R. Meir. Temporal difference based actor critic algorithms single time scale convergence and neural implementation. In *In advances in Neural Information Processing Systems*, accepted, 2008.

R. Florian. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation*, 19:1468–1502, 2007.

R. Gallager. *Discrete Stochastic Processes*. Kluwer Academic Publishers, 1995.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

H. K. Khalil. *Nonlinear Systems, 3rd Ed.* Prentice Hall, 2002.

V. Konda and V. Borkar. Actor-critic like learning algorithms for markov decision processes. 1999.

V. Konda and J. Tsitsiklis. On actor critic algorithms. *SIAM J. Control Optim.*, 42(4):1143–1166, 2003.

H. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer, 1997.

P. Marbach and J. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE. Trans. Auto. Cont.*, 46(2):191–209, 1998.

A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Annals of Applied Probability*, 16(3):1671, 2006.

J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71:1180–1190, 2008.

M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1994.

W. Schultz. Getting formal with dopamine and reward. *Neuron*, 36(2):241–63, 2002.

S. Singh and P. Dayan. Analytical mean squared error curves for temporal difference learning. *Machine Learning*, 32:5–40, 1998.

R. Sutton and A. Barto. *Reinforcement Learning*. MIT Press, 1998.

G. Tesauro. Temporal difference learning and the td-gammon. *Communication of the ACM*, 38(3), March 1995.

J. N. Tsitsiklis and B. V. Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997.